Lumigi: Shining Light on Your Process Data (Extended Abstract)

Lotte Vugs
lotte@wavespi.nl
Waves Process Intelligence
Eindhoven, The Netherlands

Maarten van Asseldonk
maarten@wavespi.nl
Waves Process Intelligence
Eindhoven, The Netherlands

Niek van Son
niek@wavespi.nl
Waves Process Intelligence
Eindhoven, The Netherlands

***Abstract*—Process mining techniques use event logs to discover process models, analyze performance, check compliance, and predict process outcomes. Since an event log is the key input for process mining, the quality of the event log is of paramount importance for the value obtained with any process mining analysis. However, data quality issues can arise while preparing event logs, e.g. inaccurate timestamps or imprecise activity names. Therefore, to ensure the insights obtained with the process mining analysis are accurate, it is important that the process data quality is validated. However, there are little structured approaches available to analyze the process data quality. In this paper, we present Lumigi. Lumigi is a freely available, stand-alone tool developed to fill the gap between the need for a structured approach to validate process data quality and the tools available for business users.**

I. Introduction

Process mining consists of a set of methods, tools and techniques to discover process models, analyze performance, check compliance and compare variants using event data recorded in information systems [1]. The key input for process mining is called an event log. In order to conduct a process mining analysis, event data needs to be collected and transformed into an event log. The quality of the event log is of utmost importance to derive maximum value from any process mining analysis. In the process of creating the event log, various data quality issues can arise, like inaccurate timestamps or imprecise activity names [2].

Although various data quality issues can reside in the event log, there are few practical guidelines for business-users to assess the quality of an event log in a structured way. Without these guidelines, a process mining analyst is forced to rely on their personal experience and trial-and-error to find data quality issues. Failing to find these issues may result in conclusions that are misleading or even flat out wrong.

This paper introduces Lumigi, a freely available, stand-alone tool that helps users to detect data quality issues for process mining through a structured approach. Lumigi is designed as a complementary tool to existing process mining tools, to support process miners in the final stages of data transformation and the first stages of the process analysis.

The remainder of this paper is structured as follows. In Section II, we describe the related work. In Section III, we outline the components of the tool. Afterwards, in Section IV contains the conclusion of the paper. Furthermore, in Section V, we outline our next steps. Lastly, in Section VI, we thank those that supported us in creating this tool.

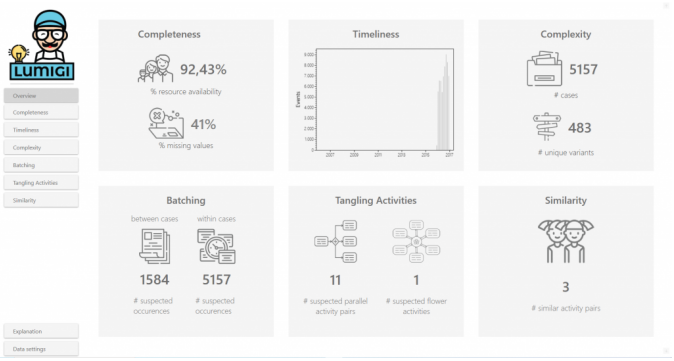

Fig. 1. Screenshot of the overview screen, depicting the different perspectives Lumigi features to analyze process data quality

II. Related Work

For a general overview of the state-of-the-art of data quality in process mining, we refer the reader to [3]. Furthermore, to the best of our knowledge, the only other tool focused on data quality assessment in process mining is the R package "DaQAPO", now part of bupaR [4].

Lumigi's key distinguishment is its focus on business users. It is based on state-of-the-art research, combined with experience from practice. To give its users a first grasp of the data quality, Lumigi features the quality dimensions *Completeness*, *Timeliness* and *Complexity*. In this context, we adopted the corresponding definitions presented in [5]. Furthermore, for *Bathcing*, *Tangling Activities*, and *Similarity*, inspiration for Lumigi is drawn from [2], [6]. In [2], the concept of *data imperfection patterns* is introduced and a list of data imperfection patterns is introduced. In [6], some of these imperfection patterns are more elaborately defined using pseudocode. Which data patterns can be found with Lumigi, and how, is outlined in Lumigi's documentation: www.lumigi.io/documentation.

III. Tool Outline

Lumigi is a freely available tool, aimed at business users. It is a stand-alone process mining tool focused on offering a structured approach to analyze process data quality. Furthermore, it is complemented with documentation explaining the

different metrics with examples, tips, and a list of possible root causes for the behavior for data quality issues found, offering a framework for enlisting possible root causes for data quality issues. A screen cast demonstrating the tool is available.[1]

A. Input Configuration

Lumigi uses a CSV file as input, after which the user is asked to specify which column represents the case identifier, the activity name, the timestamp, and, optionally, the resource.

B. Tool Overview

After configuration, all metrics are calculated, after which the user is directed to the overview screen. This screen summarizes the different perspectives that Lumigi features to analyze process data quality. A screenshot of the overview screen is shown in Figure 1. The menu can be used to navigate to a specific feature.

First, the data quality dimensions *Completeness*, *Timeliness*, and *Complexity* can be used to get a general grasp of the data. *Completeness* can be described as having all the event data that is necessary for the task at hand. This is analyzed by assessing the fraction of values in each column that is missing, and assessing the timestamp granularity. *Timeliness* measures how current the data is, and whether it is in the expected time frame. In Lumigi, this is assessed by visualizing the number of events over time. *Complexity* focuses on the structuredness of the data. On activity-level, Lumigi features, among other things, the total number of activities, the set of start activities, and the set of end activities. On variant level, Lumigi showcases the number of variants, the number of variants that occurs only once, and a plot depicting the fraction of cases per variant. The three quality dimensions that are featured by Lumigi, are designed to find relatively easy-to-find outliers before moving on to more complex data quality patterns. Examples of these easier-to-find outliers are columns with a lot of missing data, timestamp outliers, and incorrect start or end activities.

Second, more complex patterns are analyzed to find data quality issues. With *Batching*, we look at events that are recorded at (almost) the same moment in time. With his/her domain knowledge, the analyst then can reflect on whether the batching is intended, or that there is an underlying data quality issue that manifests itself through batching behavior. Here, we distinguish *between-case batching* and *within-case batching*. In academic literature, these concepts are better known as *inter-case batching* and *intra-case batching* [2]. However, we changed the terminology, as we felt it was understood better by the business users evaluating our tool.

With *Tangling Activities*, the analyst can identify activities that are suspects for tangles in process graphs, making your process discovery analysis more difficult. With this knowledge in mind, an analyst can decide to exclude some of these tangling activities in the first stage of the analysis, to structure the process graph. For this, metrics are designed to find parallel activities and activities with a lot of predecessors and successors (called *Flower Activities*).

With *Similarity*, the analyst can search for synonymous activity names.

[1]A screen cast demonstrating Lumigi can be foud here: https://www.youtube.com/watch?v=$_1eD36HlZHs$

IV. Conclusion

The quality of process data is of utmost importance to derive maximum value from a process mining analysis. However, as process miners working in practice, we found few structured approaches to analyze it. To fill this gap, this paper introduces Lumigi; a freely available, stand-alone tool focused on process data quality.

V. Next Steps

Process data quality is a developing research field. Lumigi is a first attempt to raise awareness about process data quality, but it is certainly not without flaws. We keep a list of open opportunities on our website, using the feedback we have received from our users. For the most up-to-date list of limitations of Lumigi, we refer the reader to our website: www.lumigi.io.

One of the limitations frequently mentioned is that Lumigi lacks a functionality to repair the event log. In our evaluations, business users emphasized to perform this repair as close to the source as possible. We are therefore currently expanding our focus to data transformation, and exploring ways to leverage community knowledge to create high-quality event logs.

VI. Acknowledgement

The main source of inspiration for developing Lumigi was [2]. Furthermore, the authors would like to thank the members of the Business Process Management Group of the Queensland University of Technology for their early-stage feedback, in particular: Robert Andrews, Arthur Hofstede, and Michael Adams. Furthermore, we would like to thank all process miners that provided their valuable feedback in the beta release. Although the list is too long to showcase here, and we would like to mitigate the risk of forgetting to name some of you, we hope it suffices to thank you in this fashion.

References

[1] W. Van der Aalst, *Process mining: Data science in action*, 2016.

[2] S. Suriadi, R. Andrews, A. H. ter Hofstede, and M. T. Wynn, "Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs," *Information Systems*, vol. 64, pp. 132–150, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.is.2016.07.011

[3] N. Martin, "Data Quality in Process Mining," in *Interactive Process Mining in Healthcare*, C. Fernandez-Llatas, Ed. Cham: Springer International Publishing, 2021, pp. 53–79. [Online]. Available: https://doi.org/10.1007/978-3-030-53993-1_5

[4] ——, "daqapo: Data Quality Assessment for Process-Oriented Data," 2020. [Online]. Available: https://cran.r-project.org/package=daqapo

[5] R. Verhulst, "Evaluating quality of event data within event logs: an extensible framework," *Master Thesis*, 2016.

[6] R. Andrews, M. T. Wynn, K. Vallmuur, A. H. Ter Hofstede, E. Bosley, M. Elcock, and S. Rashford, "Leveraging data quality to better prepare for process mining: An approach illustrated through analysing road trauma pre-hospital retrieval and transport processes in Queensland," *International Journal of Environmental Research and Public Health*, vol. 16, no. 7, 2019.