ElogQP: An Event log Quality Pointer

Tobias Ziolkowski[1], Lennart Brandt[2], Agnes Koschmider[1]

[1] Process Analytics Group,
Computer Science Department, Kiel University, Germany
{tzi|ak}@informatik.uni-kiel.de
[2] stu113969@mail.uni-kiel.de

Abstract. This paper presents *ElogQP*, a tool to detect data quality violations in an event log. Data quality issues significantly impact the process discovery result. Thus, *ElogQP* represents an essential step towards improved process discovery.

Keywords: event log, process mining, data cleaning, imperfection patterns.

1 Introduction

Event log files are used as input to any process mining algorithm aiming to discover an as-is process model or to identify bottlenecks. Although recently process mining has gained an impressive uptake, still, data quality violations often hamper the direct applicability of process mining techniques on an event log. There are several reasons for data quality violations like those that the recorded event data is not saved in the correct order, data entries are missing (e.g. timestamps or case ID) or are not recorded correctly (e.g. incomplete activity names). These quality violations lead to inappropriate event logs and finally significantly impact the process discovery result. To counteract data quality issues in process mining several approaches exist [1, 2, 3] like to define maturity levels for data quality [1], to use a framework of timestamp imperfections [2] or a framework for event log quality [3]. Better understanding of how data quality issues affect the event log quality led to the definition of so-called event log imperfection patterns [4].

This paper presents the Event log Quality Pointer (*EloqQP*) tool aiming to detect data quality violations. The tool allows to detect event log imperfection patterns and to classify the data violations according to data quality levels as specified in the process mining manifesto [5]. Beside this, a comparison between two event logs with respect to data quality violations is supported. Thus, *ElogQP* detects missing start or end activities and activities with incorrect order. **Fig. 1** shows how *ElogQP* works when two event logs are used as input. The event log on the left-hand side is (more) complete, while on the right-hand side one timestamp and one activity are missing. When parsing both event logs, *ElogQP* returns data types that have been identified as data quality violations with a descriptive comment to understand the violation (see "*Output of ElogQP*").

The paper is structured as follows. Section 2 gives an overview of *ElogQP*. It describes the components and the functionality of the tool. Section 3 concludes the paper.

J. Manner, S. Haarmann, S. Kolb, N. Herzberg, O. Kopp (Eds.): 13th ZEUS Workshop, ZEUS 2021, Bamberg, held virtually due to Covid-19 pandemic, Germany, 25-26 February 2021, published at http://ceur-ws.org

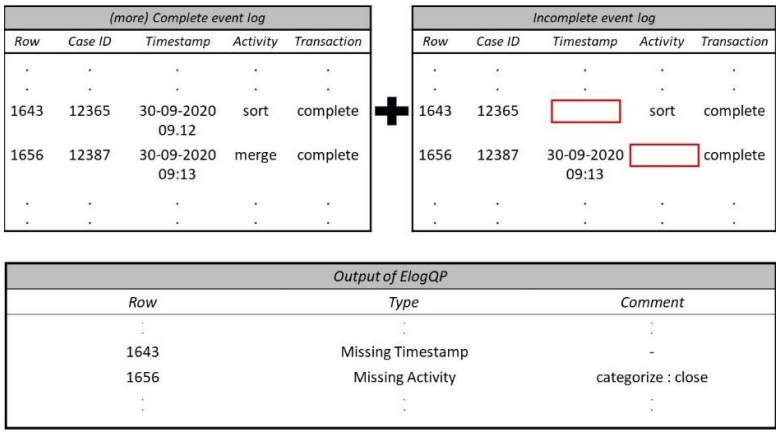

Fig. 1. *ElogQP* detects missing timestamp and missing activity.

2 Detection of Data Quality Violations

The next section summarizes event log imperfection patterns and data quality levels of an event log. Section 2.2. presents how *ElogQP* refers to both.

2.1 Event Log Imperfection Patterns and Data Quality Levels

Eleven event log imperfection patterns for process mining have been defined, which are form-based event capture, inadvertent time travel, scattered event, elusive case, scattered case, collateral events, polluted label, distorted label, synonymous labels and homonymous labels. These patterns relate to data quality issues in timestamps, case IDs, activities and activity labels like missing or incorrect activities, missing case IDs and discrepancies in the activity names.
According to the process mining manifesto [5] five quality levels exist for event logs. Quality level 1 means that the recorded events do not exist in reality and thus the event log has artificial events. Often these are manually created event logs. Quality level 2 refers to event logs that are recorded without a systematic approach. This returns log data that is incorrect or incomplete. Event logs with a quality level 3 are reliable in a way that the recorded event data is likely to correspond with reality. Quality level 4 means that event logs are complete in terms of "correct". Quality level 5 fulfills the properties of quality level 4. Additionally, the recorded events have clear semantics and are well defined. *ElogQP* evaluates quality violations according to quality level 1 to 4.

2.2 Tool Overview

Fig. 2 shows the functionality of the *ElogQP* tool. The tool has been implemented in R and in essence, the tool represents a script with the following sequential steps:

- *Step 1*: The event log is imported in XES format into the *ElogQP* environment.
- *Step 2(a)*: The user selects the event log quality attributes to be analyzed.
- *Step 2(b)*: An additional event log or Petri net can be used as input. The comparison between the Petri net and an event log additionally allows detecting activity order incompliance. With the additional event log missing attributes can be detected.
- *Step 3*: The event log is analyzed according to the selected attributes.
- *Output*: If any data quality issue is found, *ElogQP* sets a pointer, indicates the data quality level and returns a descriptive comment as shown exemplary in Fig. 1.

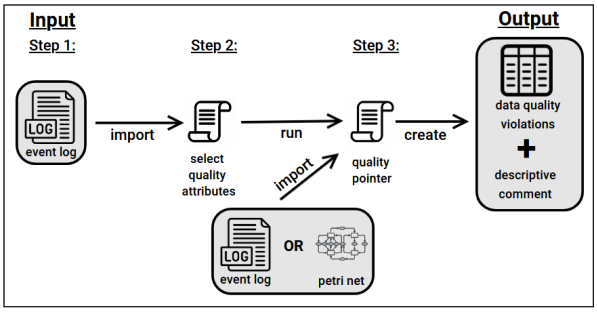

Fig. 2. How *ElogQP* works

Fig. 3 shows the output of *ElogQP* with a quality level of 2 and the detected data quality violations. If no data quality violations are found, a quality level of 4 is returned.

Row		Typ	Comment
		Event log with quality 2	
1430	6	Missing Activity	check vacation area : Determine budget
1431	7	Missing Activity	Determine budget : determine weather
1432	8	Missing Activity	determine weather : Determine holiday type
1433	9	Missing Activity	Determine holiday type : Ask employer for vacation
1434	10	Missing Activity	Ask employer for vacation : Determine desired destination
1435	11	Missing Activity	Determine desired destination : Ask employer for vacation
1436	13	Missing Activity	Determine duration : Send offers
1437	14	Missing Activity	Send offers : Check form
1	15	Wrong Name	Check form
1438	15	Missing Activity	Check form : Wait for corrected form
2	16	Wrong Name	Wait for corrected form
1439	16	Missing Activity	Wait for corrected form : Check form
1440	17	Missing Activity	Check form : Wait for corrected form
1441	18	Missing Activity	Wait for corrected form : Logout
1443	19	Missing End	Logout
1442	20	Missing Start	register
1444	23	Missing Activity	Determine budget : check vacation area
1445	24	Missing Activity	check vacation area : Check criteria

Fig. 3. Interface of *ElogQP*

3 Conclusion and Future Work

This paper presented *ElogQP*, a tool to inspect event logs to find data quality violations. In this way, *ElogQP* is a tool for cleaning event logs thus improving the process discovery result. In future work we plan to completely implement all event log imperfection patterns. So far, *ElogQP* does not detect unanchored events, elusive case and scattered case. Additionally, we will integrate data quality recommendations that have been suggested for process activity labels [6] into *EloqQP*.

References

1. Leemans, M., van der Aalst, W.M.P.: Discovery of frequent episodes in event logs. In: SIMPDA 2014: 31-45, vol. 1293 of CEUR Workshop Proceedings
2. Fischer, D. A., Goel, K., Andrews, R., Dun, C. G. J. van, Wynn, M.T., Röglinger, M.: Enhancing Event Log Quality: Detecting and Quantifying Timestamp Imperfections. BPM 2020, vol. 12168 of LNCS, Springer, pp. 309-326.
3. Kherbouche, O. M., Laga, N., Masse, P.-A. (2016): Towards a better assessment of event logs quality. SSCI 2016, IEEE, pp. 1-8.
4. Suriadi, S., Andrews, R., Hofstede, A.H.M. ter, Wynn, M.T. (2017): Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs. Information Systems 64: 132-150: https://doi.org/10.1016/j.is.2016.07.011.
5. van der Aalst, W.M.P. et al. (2012) Process Mining Manifesto. Business Process Management Workshops (1) 2011: 169-194, https://doi.org/10.1007/978-3-642-28108-2_19.
6. Koschmider, A., Ullrich, M., Heine, A., Oberweis, A. (2015): Revising the Vocabulary of Business Process Element Labels. CAiSE 2015, vol. 9097 of LNCS, Springer, pp. 69-8.