Responsible Process Mining - A Data Quality Perspective

Moe Thandar Wynn[1(✉)] and Shazia Sadiq[2]

[1] Queensland University of Technology, Brisbane, Australia
m.wynn@qut.edu.au
[2] The University of Queensland, Brisbane, Australia
shazia@itee.uq.edu.au

Abstract. Modern organisations consider data to be their lifeblood. The potential benefits of data-driven analyses include a better understanding of business performance and more-informed decision making for business growth. A key road block to this vision is the lack of transparency surrounding the quality of data. A process mining study that utilises low-quality, unrepresentative data as input has little or no value for the organisation and becomes a catalyst for erroneous conclusions ('Garbage-in-Garbage-out'). Many process mining techniques do not take into account inherent inaccuracies in the data, or how the data might have been manipulated or pre-processed. It is thus impossible to ascertain the degree to which analysis outcomes can be relied upon. This tutorial paper outlines foundational concepts of data quality with a special focus on typical data quality issues found in event data used for process mining analyses. Key challenges and possible approaches to tackle these data quality problems are elaborated on.

Keywords: Process mining · Data quality

1 Introduction

Process Mining is a specialised form of data-driven process analytics where data about process executions, collated from the different IT systems typically available in organisations, is analysed to uncover *the real behaviour and performance* of business operations [2]. Without question, the extent to which the outcomes from process mining analyses can be relied upon for insights is directly related to the quality of the input data. The onus is usually on a process analyst to identify, assess and appropriately remedy data quality issues so as to avoid inadvertently introducing errors into the data while minimising information loss. It is widely acknowledged that eighty percent of the work of data scientists is taken up by data preparation and handling data quality issues[1]. The case of the process analyst is no different [17].

[1] https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#58f51e5d6f63.

T. Hildebrandt et al. (Eds.): BPM 2019, LNCS 11675, pp. 10–15, 2019.
https://doi.org/10.1007/978-3-030-26619-6_2

There has been an increased interest in research investigating the issues of responsible data science [3,15]. Key dimensions in the notion of responsible data science (such as fairness, accuracy, transparency, and confidentiality [3]) are being explored and also for different domains (e.g., healthcare). In order to take steps towards responsible process mining, there is the dual need to increase the importance of data quality awareness and mitigate the opportunity to make erroneous conclusions, while helping process analysts overcome the burden of managing data quality.

In this tutorial paper, we focus on event logs as the primary form of input into process mining. Accordingly, we first present a brief summary of existing work on understanding data quality requirements for event logs. In the words of Edward Demming, the father of quality management, you can't manage what you can't measure. Hence, our next section outlines key techniques for measuring data quality in event logs. Finally, we provide a synopsis of current contributions and future needs of data quality awareness in process mining.

2 Understanding Data Quality Requirements for Event Logs

An event log used for process mining contains a collection of cases whereby each case can be seen as a sequence of events [2]. Each event refers to a case, an activity being undertaken, a point in time and a transaction type. An event may also refer to a resource or an organisational role and other data attributes (e.g., customer details and case outcomes).

The process mining manifesto [1] highlighted the need for high-quality event logs for process mining. The manifesto describes five levels of maturity ranging from one star to five stars. At the lowest level of maturity (*) where events are recorded manually, one may find that events that are incorrectly entered (e.g., incorrect timestamps or activity labels) or events may be missing. At the highest level of maturity (*****), event logs are considered to be complete and accurate as events are recorded automatically by a system (e.g., a process-aware information system). Most real-life event logs are found to be in-between these two extremes of the scale with many quality issues [6,17].

As most process mining techniques make use of key event data, namely, case identifiers, activity labels, and timestamps, missing, inaccurate or erroneous values (e.g., only a date is recorded but no time, incorrect spellings or variations in how activities are labeled) for any of this data may mean that a case or an event has to be filtered out or an erroneous value may need to be replaced, or a missing value may need to be inferred.

Given the diversity of data quality problems, it is important to understand the key requirements. While Juran and Godfrey [10] provide the fundamental "fitness for use" principle, decades of data quality research has proliferated various understandings of data quality requirements through its underlying dimensions [8,14,16,20]. Over the course of time, many of the definitions for different data quality dimensions have overlapped, and the same definitions for the same

dimensions have developed conflicting interpretations, resulting in a level of disparity that does not support a shared understanding. Recent work offers an empirically validated consolidation of these dimensions covering both academic and practitioner perspectives [9], and provides 33 dimensions clustered into eight categories, namely Completeness, Accuracy, Validity, Consistency, Currency, Availability and Accessibility, Reliability and Credibility, and Usability and Interpretability. These studies indicate that data quality requirements cover both objective (e.g. uniqueness and format consistency) as well as subjective (e.g. relevance and freshness) dimensions.

There have been efforts by process mining researchers to classify data quality issues typically found in event logs [6,12,17,18] with a view to take steps towards addressing these issues and thus to increase the reliability of analysis results.

Bose et al. [6] identify four broad categories of issues affecting event log quality: missing data (where data items are not recorded in an event log), incorrect data (where data items are incorrectly recorded in an event log), imprecise data (where recorded values are considered too coarse to be useful) and irrelevant data (where data items contains irrelevant information). The authors also identify 27 classes of event log quality issues (e.g., problems related to timestamps in event logs, imprecise activity names, and missing events) depending on where they occur such as cases, events, activity labels, timestamps, resources. Their intention is to "encourage systematic logging approaches (to prevent event log issues), repair techniques (to alleviate event log issues) and analysis techniques (to deal with the manifestation of process characteristics in event logs)" [6]. These issues were illustrated from the analysis of five real-life event logs from different application domains.

Suriadi et al. [17] identify eleven event log imperfection patterns based on their experience with over 20 Australian industry data sets which confirm the severity of data quality issues in process data and their potential impact on process mining analyses. The eleven patterns include form-based event capture, inadvertent time travel, unanchored event, scattered event, elusive case, scattered case, collateral event, polluted label, distorted label, synonymous labels and homonymous label. Each pattern is described using the following components: description of the pattern, real-life example of the pattern, affect which captures the consequence of the occurrence of the pattern on process mining outcomes, the type of data event and event log entities affected by the pattern, strategy to detect the presence of a pattern, potential remedies and side-effects of these remedies, and indicative rules for detection.

Lu and Fahland [12] propose a conceptual framework to better understand event data quality for process mining analysis. The framework categorises event data into three entities: quality of events, quality of ordering of events and quality of labels of event. These three entities are then evaluated based on two dimensions: individual trustworthiness and global conclusiveness whereby individual trustworthiness focuses on the intrinsic qualities of event data (e.g., accuracy or correctness dimensions) while the global conclusiveness indicates if a significant pattern is being observed.

3 Measuring Data Quality of Event Logs

Data quality requirements continue to be dictated by the fitness for use principle [10], thus making them highly dependent on the use context. Further a plethora of diversified requirements (i.e. dimensions) exist, which are in turn deeply bound to use context making them complicated to model, analyse, and re-use, resulting in a prohibitive capacity to have a common set of measures for detecting and quantifying data quality.

Batini et al. [4] provide a comprehensive analysis of existing approaches for data quality assessment. We note that most, if not all, of these approaches follow a user centric approach where requirements are solicited from users before the data is explored (see e.g. [5,11,19]).

However, in the process mining context, access to the creators of data that constitutes event-logs cannot be relied on. This is mostly the case for publicly available event logs. Furthermore a process analyst cannot typically influence data capture practices and hence expectation of cleaning of the source data may be misplaced. Thus it is imperative to measure the quality of an event log respective to the particular type of analysis intended such as process discovery, performance analysis or conformance checking. For instance, the missing values metric assesses the fraction of the log for which a particular log attribute is populated which contributes to quantifying the Completeness dimension. In a log where the majority of events only have "complete" (rather than "start" and "complete") timestamps, i.e. have a high degree of missing values, the suitability of that log for performance analysis is low while the suitability for process discovery may not be negatively affected. On the other hand, if recorded timestamps do not accurately reflect when an activity occurred, process discovery will be compromised.

In [18], the authors propose an extensible framework to measure event data quality based on twelve dimensions collated from prior literature and to quantify the prevalence of data quality issues in event data. They include completeness, uniqueness, timeliness, validity, accuracy/correctness, consistency, believability, credibility, relevancy, security/confidentiality, complexity, coherence, representation/format.

Another early advocate of detecting data quality issues in event logs is Anna Rozinat, the co-founder of Disco Process Mining Tool. Through a number of blog posts which have now been collated into a book on process mining in practice[2], various data quality issues in event logs and ways to detect and (potentially) repair them were discussed. The quality issues mentioned in the book include formatting errors, missing data (event, attribute values, case IDs, activities, timestamps, attribute history, timestamps for activity repetition) as well as zero timestamps, wrong timestamps, same timestamps for multiple activities and different timestamp granularity.

[2] http://processminingbook.com.

4 Data Quality Awareness in Process Mining

Keeping a detailed record of the origins of data and how data is transformed along the way will increase its traceability and trustworthiness. Where such information is unavailable, the extent and effect of changes on the data will be opaque to the analyst who, may view the data as 'ground truth', i.e. direct observations as opposed to already modified data. Such a view can result in inaccurate or misleading analysis results or inappropriate further transformations. For instance, where the analyst is unaware that a data set extracted from a hospital's emergency department has been modified through time-shifting in order to de-identify patients (as in the case of MIMIIC critical care data set[3]), using this data for performance analysis will lead to incorrect results.

There has been some work to detect and repair quality issues associated with event logs. For example, Dixit et al. [7] presents a user-guided technique to detect event ordering imperfection patterns in a log associated with timestamps and then repairing identified issues using user input. The timestamp related quality issues such as different granularities, order anomaly and statistical anomaly are detected and repaired. Similarly, Lu et al. [13] presents an interactive way to assist users explore data quality patterns of interest using the context information contained in an event log. Five measures to quantify the pervasiveness of a pattern in an event log are also proposed. They include the pattern support, pattern confidence, case support, case confidence and case coverage.

To date there has been little research aimed at developing a comprehensive framework to address the issue of incorrect analysis results from inadequate data quality of event logs. Lessons from prior work in quality awareness for database (e.g., [21]) indicate that there are at least three essential components of such frameworks, each of which presents a number of research challenges, namely (1) data quality profiling that builds on shared understanding of data quality dimensions and associated metrics, (2) user preference modelling that allows users analytic needs to be captured, and (3) visibility of quality profiles together with analysis (process mining) results to improve understanding of the impact of inadequate data quality. We invite process mining researchers to tackle these challenges to move towards responsible process mining with the aim to improve the credibility and trust of stakeholders in process mining results.

Acknowledgements. The authors would like to acknowledge the input from QUT researchers (Professor ter Hofstede, Dr Andrews, Dr Suriadi and Dr Poppe) who work on this topic. This work is partly supported by ARC Discovery Project DP190102141 on Building Crowd Sourced Data Curation Processes.

References

1. van der Aalst, W., et al.: Process mining manifesto. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) BPM 2011. LNBIP, vol. 99, pp. 169–194. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28108-2_19

[3] https://mimic.physionet.org/.

2. Van der Aalst, W.M.P.: Process Mining: Data Science in Action, 2nd edn. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-662-49851-4d
3. van der Aalst, W.M.P., Bichler, M., Heinzl, A.: Responsible data science. Bus. Inf. Syst. Eng. **59**(5), 311–313 (2017)
4. Batini, C., Cappiello, C., Francalanci, C., Maurino, A.: Methodologies for data quality assessment and improvement. ACM Comput. Surv. **41**(3), 16 (2009)
5. Batini, C., Scannapieco, M.: Data Quality: Concepts, Methodologies and Techniques. Springer, Heidelberg (2006). https://doi.org/10.1007/3-540-33173-5
6. Bose, J.C., Mans, R., van der Aalst, W.M.P.: Wanna improve process mining results - it's high time we consider data quality issues seriously. In: IEEE Symposium on Computational Intelligence and Data Mining, pp. 127–134 (2013)
7. Dixit, P.M., et al.: Detection and interactive repair of event ordering imperfection in process logs. In: Krogstie, J., Reijers, H.A. (eds.) CAiSE 2018. LNCS, vol. 10816, pp. 274–290. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91563-0_17
8. Eppler, M.J.: Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes. Springer, Heidelberg (2006). https://doi.org/10.1007/3-540-32225-6
9. Jayawardene, V., Sadiq, S., Indulska, M.: The curse of dimensionality in data quality. In: 24th Australasian Conference on Information Systems (ACIS), pp. 1–12. RMIT University (2013)
10. Juran, J., Godfrey, A.: Quality Handbook. Republished McGraw-Hill, New York (1999)
11. Lee, Y.W., Strong, D.M., Kahn, B.K., Wang, R.Y.: AIMQ: a methodology for information quality assessment. Inf. Manag. **40**(2), 133–146 (2002)
12. Lu, X., Fahland, D.: A conceptual framework for understanding event data quality for behavior analysis. In: Kopp, O., Lenhard, J., Pautasso, C. (eds.) Central European Workshop on Services and their Composition ZEUS. CEUR Workshop Proceeedings, vol. 1826, pp. 11–14 (2017)
13. Lu, X., et al.: Semi-supervised log pattern detection and exploration using event concurrence and contextual information. In: Panetto, H., et al. (eds.) OTM 2017. LNCS, vol. 10573, pp. 154–174. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69462-7_11
14. Scannapieco, M., Catarci, T.: Data quality under a computer science perspective. Arch. Comput. **2**, 1–15 (2002)
15. Srivastava, D., Scannapieco, M., Redman, T.C.: Ensuring high-quality private data for responsible data science: vision and challenges. J. Data Inf. Qual. **11**(1), 1:1–1:9 (2019)
16. Stvilia, B., Gasser, L., Twidale, M.B., Smith, L.C.: A framework for information quality assessment. J. Am. Soc. Inform. Sci. Technol. **58**(12), 1720–1733 (2007)
17. Suriadi, S., Andrews, R., ter Hofstede, A.H.M., Wynn, M.T.: Event log imperfection patterns for process mining: towards a systematic approach to cleaning event logs. Inf. Syst. **64**, 132–150 (2017)
18. Verhulst, R.: Evaluating quality of event data within event logs: an extensible framework. Master's thesis, Technische Universiteit Eindhoven, August 2016
19. Wang, R.Y.: A product perspective on total data quality management. Commun. ACM **41**(2), 58–65 (1998)
20. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. J. Manag. Inf. Syst. **12**(4), 5–33 (1996)
21. Yeganeh, N.K., Sadiq, S., Sharaf, M.A.: A framework for data quality aware query systems. Inf. Syst. **46**, 24–44 (2014)