A Conceptual Framework for Understanding Event Data Quality in Behavior Analysis

Xixi Lu and Dirk Fahland

Eindhoven University of Technology, The Netherlands
{x.lu,d.fahland}@tue.nl

1 Background and Motivation

Process mining aims to derive useful insight for improving business process efficiency and effectiveness. These mining techniques rely heavily on event data, in the form of *event logs*, to provide accurate diagnostic information. The quality of such event data therefore has a large effect on the quality and trustworthiness of the conclusions drawn from the mining analysis and the subsequent business decisions made.

Traditional data quality frameworks focus on identifying *quality dimensions* extensively from a *data perspective* and improving the overall data quality in the long term. While long-term data quality improvement is certainly useful, this may not aid analysts in practice who are often faced with the task of analyzing a given log of lower quality in the short term. As result, when the user conducts a certain analysis (e.g., process discovery), these quality frameworks provide little guidance for assessing or improving the quality of data for the analysis [1, 2, 7].

To the best of our knowledge, only the work in [7] presented event data quality issues as specific patterns reoccurring in logs and discussed their possible effects on mining results from an *analysis perspective*.

In the past few years, we have developed numerous approaches to deal with event logs of low quality, for which no conclusive results are obtained when the user applies existing mining techniques. Three main approaches have emerged: (i) a trace clustering technique based on behavior similarity which allows the user to identify process variants and then explore these variants to discover more precise and conclusive models [4]; (ii) a conformance checking technique using partial order traces and alignments should the ordering of events in a log be untrustworthy [3]; (iii) a label refinement technique in cases where labels of events are imprecise and lead to inconclusive models [5]. However, as each approach is dedicated to tackle a particular event data quality issue from an analysis perspective, an overview for understanding the quality issues is missing.

In this positioning paper, we would like to discuss a conceptual framework to help users understand how these quality issues could be presented and interrelated, how our approaches may be positioned and how future data quality issues may be classified. The conceptual framework[1] is visualized as a table: the columns

[1] The term conceptual framework has taken different definitions in different contexts [6, Chap. 1]. In this paper, we consider a conceptual framework as an analytic tool that helps the user to understand and distinguish different concepts and is easy to remember and apply.

O. Kopp, J. Lenhard, C. Pautasso (Eds.): 9th ZEUS Workshop, ZEUS 2017, Lugano, Switzerland, 13-14 February 2017, published at http://ceur-ws.org/

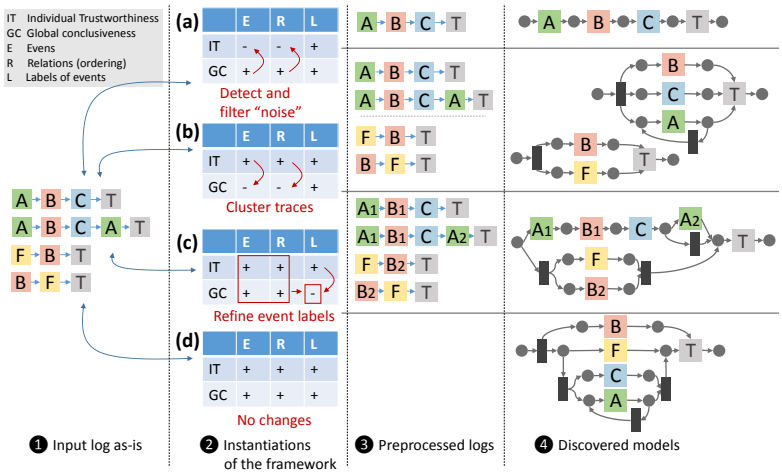

Fig. 1: Four examples of quality issues visualized as possible instantiations of the framework, and the possible preprocessing steps followed.

outline the entities in input data (logs or models) that are relevant for behavior (control-flow) focused analysis; the rows list two dimensions of quality, *individual trustworthiness (IT)* and *global conclusiveness (GC)* which assess the quality of event data from a data perspective and an analysis perspective, respectively. Figure 1 shows four instantiations of the conceptual framework for an event log and is discussed more in depth in Section 2.

2 The Conceptual Framework

In this section, we first explain the framework, its columns, rows and the values assigned to each cell. Secondly, we discuss four prominent cases of event data quality issues and how they are captured by the framework. Finally, we discuss how to extend the framework to capture other cases and conclude the paper.

Our studies into event data quality have shown that there are three entities in event logs, whose quality or trustworthiness have an effect on the results of behavior analysis (e.g., process discovery or compliance checking): (1) quality of *events (E)*, (2) quality of ordering of events, or *relations* among events *(R)*, and (3) quality of *labels* of events *(L)*. These three entities therefore constitute the columns in the framework.

The quality of each entity is divided into two dimensions: *individual trustworthiness* and *global conclusiveness*. Individual trustworthiness expresses all intrinsic qualities of event data; basically the trust of the user regarding how accurate the event data reflects the real process executions. This quality dimension is similar to *accuracy* or *correctness* dimensions discussed in the literature [1]. However, little research has been conducted into measuring this quality dimension of event data sets. We propose to have three possible values for the individual trustworthiness,

as the aim is to allow the user to use the framework with ease and obtain a quick impression of the quality of the data. The three value includes: +, which indicates that the user assumes 100% trustworthiness; −, which refers to that there are some non-trustworthiness but the majority are trustworthy; −−, which refers to largely untrustworthy data. For example, the user assigns the individual quality of *events* of a log a + if all events fully reflect the process execution (e.g., fully automated recordings); the user may assign a − if the user thinks there are a few events missing or some duplicated (e.g., when two doctors attended the same consultation for a patient, the consultation might be recorded as two events for the same patient). As another example, the ordering of events in a log might be assigned with − should the user observe that many events happened on the same date and no time is recorded.

Global conclusiveness indicates whether there is a certain path, a certain structure, or a certain pattern that can be observed and is significant, indicating such a pattern is not a random artifact. In other words, this dimension assesses whether there is some behavior, possibly unknown, shared and repeated across a significant number of cases, which implies that there is a particular mechanism or force controlling the flow. Having such a mechanism indicates that future cases would most likely follow this mechanism or pattern. We assign a + if such mechanism is significant enough in a log to be observed and concluded, otherwise a minus −. The lack of conclusiveness might indicate the behavior is random or unique, rendering the results of process analysis useless. We acknowledge that conclusiveness is rather difficult to assess or to attribute to only the log or only the model, because conclusiveness may also depend both on the technique applied and on the expectations or understanding of the user of the results. The user may therefore reassess conclusiveness based on the results obtained. Note that there is no trade-off between the two dimensions, a good event data set should be both trustworthy and conclusive in order to perform analysis.

Examples. Figure 1 exemplifies four cases of the framework: the log as-is is shown on the left-hand side; the four tables, one for each case, are shown in the middle of Figure 1; the preprocessed logs and corresponding models are shown on the right-hand side. The first case (a) in Figure 1 is well-known: the user classifies the log as containing some non-trustworthy events (and relations), thus '−' for IT of the events and relations. Nevertheless, the log shows the normative behavior (main-flow) rather conclusively, thus '+' for GC. Then the analyst may tackle this issue by removing the non-trustworthy cases (or events) and discovering a model from the trustworthy cases. Assuming that the variant $\langle A, B, C, T\rangle$ is very frequent and concludes the main behavior, the user can filter out the other cases and discover a simple, sequential model based on this variant.

In contrast, one might classify the same log as globally inconclusive ('−' for GC of the events and relations) but individual event as trustworthy ('+' for IT), then we have case (b) in Figure 1. To improve the conclusiveness, one might cluster the traces using behavior similarity, since the events and relations among the events are classified as trustworthy. For each cluster, a more precise and conclusive model may be discovered [4]. As in the third case (c) in Figure 1,

the user considers some event labels as inconclusive ('−' for GC of the labels), while the rest of the log is classified as trustworthy and should contain the main behavior ('+' for the rest). Then one may use this information to refine the labels of events by finding similar groups of events that share the same label and the same context. The log with refined labels then yields more conclusive results [5]. Finally, in case of (d) in which all entities are classified as trustworthy and conclusive, one may simply discover a model. If the resulting model is of low quality or inconclusive, this is an indication that the quality of the result does not reflect the (previously assessed) quality of the input. One might revisit the table, reconsider the values assigned to each cell (especially conclusiveness) regarding the applied analysis technique, and improve the quality by preprocessing the log.

Outlook. The columns and the rows of the framework could be extended to tailor the framework towards other analyses. For example, if the user conducts performance analyses in addition to behavior analysis, one could add the entity *timestamps* as a fourth column. Similarly, resources or data attributes could be added as columns. The rows could be extended to other quality dimensions of importance. Interestingly, one may add *model quality* as a row, assessing the quality of the model as an additional quality dimension in the context of a conformance checking analysis. As future work, the applicability of the conceptual framework may be evaluated by conducting empirical studies involving process experts or analysts.

References

1. Bose, R.P.J.C., Mans, R.S., van der Aalst, W.M.P.: Wanna improve process mining results? In: Computational Intelligence and Data Mining (CIDM), 2013 IEEE Symposium on. pp. 127–134. IEEE (2013)
2. Gschwandtner, T., Gärtner, J., Aigner, W., Miksch, S.: A taxonomy of dirty time-oriented data. In: International Conference on Availability, Reliability, and Security. pp. 58–72 (2012), http://dx.doi.org/10.1007/978-3-642-32498-7_5
3. Lu, X., Fahland, D., van der Aalst, W.M.P.: Conformance checking based on partially ordered event data. In: Business Process Management Workshops - BPM 2014 International Workshops, Eindhoven, The Netherlands, September 7-8, 2014, Revised Papers. pp. 75–88 (2014), http://dx.doi.org/10.1007/978-3-319-15895-2_7
4. Lu, X., Fahland, D., van den Biggelaar, F.J.H.M., van der Aalst, W.M.P.: Detecting deviating behaviors without models. In: BPM Workshops 2015. pp. 126–139. Springer (2015)
5. Lu, X., Fahland, D., van den Biggelaar, F.J.H.M., van der Aalst, W.M.P.: Handling duplicated tasks in process discovery by refining event labels. In: BPM 2016. pp. 90–107. Springer (2016)
6. Ravitch, S.M., Riggan, M.: Reason & rigor: How conceptual frameworks guide research. Sage Publications (2016)
7. 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), http://dx.doi.org/10.1016/j.is.2016.07.011