Estimating Process Conformance by Trace Sampling and Result Approximation (Extended Abstract)[1]

Martin Bauer, Han van der Aa, Matthias Weidlich[2]

1 Motivation

Process-oriented information systems allow for the coordination of process models, describing the interplay between activities drawn out to reach a certain business goal, and recorded data, which captures the execution of these activities in real life. In particular, conformance checking methods, allow for the analysis and the comparison of these two views on a business process at hand, thus enabling a process analyst to comprehensively answer the question of how envisioned and executed behavior relate to each other [Ca18]. As the volume and frequency with which data is recorded increases, event logs comprise billions of events [VdA14]. Due to the exponential run time complexity of alignment-based conformance checking techniques [VD18], which are considered the de facto standard, the need for run time improvements is apparent. While various techniques for efficient alignment computation have been proposed, these, fundamentally, still require the consideration of *all*, possibly billions, of recorded events in a log.

Often, however, conformance checking aims to provide a general understanding of the overall conformance of process execution, rather then identifying all individual deviations. Recognizing this, we argue that for the calculation of general conformance insights, only a representative fraction of the traces in a log is required. Therefore, we are able to improve the run time performance of conformance checking in two ways. First, we reduce the number of traces required to obtain a conformance result through a sampling procedure. Second, we further reduce the required number of alignment calculations by using a worst-case approximation of the conformance result, for traces that are sufficiently similar to previously seen ones. In our work, we instantiate both the sampling and the approximation methods for two types of conformance results: a numerical fitness measure and a distribution of conformance issues over all activities.

[1] The original article was presented at the 17th International Conference on Business Process Management (2019) in Vienna, Austria [BVdAW19]

[2] All authors are with the Humbold-Universität zu Berlin, Department of Computer Science, Berlin, Germany, {martin.bauer | han.van.der.aa | matthias.weilich}@hu-berlin.de

2 Approach

For trace sampling, we formulate the sampling of a new trace, calculating its conformance result, and updating the aggregated conformance result as a series of binomial experiments. In particular, by quantifying change induced in the aggregated conformance result ϵ, we can view this procedure as a binomial trial with two outcomes: either the trace introduces new information or it does not. Furthermore, based on a statistically grounded minimal sample size N, we determine when enough traces have been seen to consider the conformance result as sufficiently representative of the complete log. Formally, N represents the number of consecutive traces without new information that are required to conclude that the unknown probability p of the next trace containing new information is less than a chosen parameter δ.

To approximate conformance results, we compute the maximal impact that a newly sampled trace ξ can have on an aggregated conformance result. This approximation is based on the most similar observed trace ξ', as well as its edit distance to ξ. The worst-case impact of ξ on the overall conformance result is approximated based on the alignment of ξ' plus all differences between ξ and ξ'. The potential impact of ξ on the aggregated conformance result is checked against a significance threshold ϵ. Only if this check signals a potential significant change, an alignment is computed for ξ. Otherwise, we instead use the estimation as the conformance result of ξ, thus avoiding the need to compute an alignment for ξ.

3 Evaluation Results

We evaluated our techniques in comprehensive experiments with real-world and synthetic datasets for conformance results resembling the global fitness of the log and the relative distributions of non-conformant activities. Our results highlight dramatic improvements in terms of conformance checking efficiency compared to the baseline approach [VD18]. With samples as small as 0.1% to 1% of a log, we obtain conformance results that are virtually equivalent to those obtained when considering the complete log. This translates into respective reductions of the run times of conformance checking algorithms.

References

[BVdAW19] Bauer, Martin; Van der Aa, Han; Weidlich, Matthias: Estimating Process Conformance by Trace Sampling and Result Approximation. In: BPM. pp. 179–197, 2019.

[Ca18] Carmona, Josep; Van Dongen, Boudewijn F.; Solti, Adreas; Weidlich, Matthias: Conformance Checking – Relating Processes and Models. Springer, 2018.

[VD18] Van Dongen, Boudewijn F.: Efficiently Computing Alignments - Using the Extended Marking Equation. In: BPM. pp. 197–214, 2018.

[VdA14] Van der Aalst, Wil M. P.: Data Scientist: The Engineer of the Future. Springer International Publishing, Cham, pp. 13–26, 2014.