Activity and Sequence Detection Evaluation Metrics: A Comprehensive Tool for Event Log Comparison

Aaron Friedrich Kurz[1,*], Ronny Seiger[1], Marco Franceschetti[1] and Barbara Weber[1]

[1]University of St. Gallen, Switzerland

Abstract

Nowadays, event logs are not only created by traditional information systems, but also new data sources such as the IoT are considered to derive and construct event logs. This makes it necessary to evaluate the quality of these detected event logs and their underlying detection methods by comparison with given ground truth logs. We present *AquDeM*, enabling the comparison of XES-based event logs to evaluate activity and sequence detection methods. AquDeM features 1) a Python library that allows for programmatic comparison of event logs featuring a comprehensive set of metrics, and 2) a web app for visual event log comparison.

Keywords

Business Process Management, Internet of Things, Activity Recognition, Activity Detection, Sequence Detection, Event Log Comparison

1. Introduction

An often investigated subject at the intersection of Business Process Management (BPM) and the Internet of Things (IoT) is the abstraction of low-level IoT events to BPM-level activities [1], which can be seen as a multi-class activity detection problem. In previous work, we presented a corresponding method [2] that has the goal of detecting business process activities in real-time, based on annotated IoT data streams to enable online process conformance checking [3]. While investigating methods to evaluate the quality of the IoT-based detection of activities, which we capture in event logs in XES [4] format, we realized that most event log comparison tools that exist in the BPM field are not suitable to evaluate activity detection methods. They do not provide helpful metrics for the comparison of a *detected* event log (e.g., created from IoT data by the detection method) with a *ground truth* event log (representing the correct sequence and timing of activities as manually annotated or predefined) to evaluate and improve a specific detection method, since they *i)* focus on variant comparison to derive insights for process analysts regarding business outcomes (i.e., comparing process performance indicators) [5]; and/or *ii)* they produce results that are not suitable for rapid comparison due to non-quantitative outputs

Proceedings of the Best BPM Dissertation Award, Doctoral Consortium, and Demonstrations & Resources Forum co-located with 22nd International Conference on Business Process Management (BPM 2024), Krakow, Poland, September 1st to 6th, 2024.

*Corresponding author.

aaron.kurz@unisg.ch (A. F. Kurz); ronny.seiger@unisg.ch (R. Seiger); marco.franceschetti@unisg.ch (M. Franceschetti); barbara.weber@unisg.ch (B. Weber)

0000-0002-2547-6780 (A. F. Kurz); 0000-0003-1675-2592 (R. Seiger); 0000-0001-7030-282X (M. Franceschetti); 0000-0002-6004-4860 (B. Weber)

(e.g., graphs or natural language) [5]. Thus, we sought out methods from other relevant fields in the literature.

Core **requirements** for the event log comparison tools and metrics derived from our use-cases [3] are: *i)* they need to provide insights relevant for the detection quality of an activity detection method, i.e., on whether the activities are detected to be in-/active at the right times and/or whether the sequence of detected activities is correct w.r.t. a given ground truth; *ii)* they need to provide quantitative results for rapid and automatable comparison (e.g., for programmatic exploration of a potentially large number of method parameters); and *iii)* they should provide insights over multiple cases within the event logs. We found multiple suitable metrics in the areas of information theory, signal processing, and general activity recognition. However, none of them could be directly applied to BPM-related concepts (i.e., XES-based event logs): some important metrics (e.g., from [6]) were not available as open implementations, and some others needed modification to make sense in the context of BPM (e.g., cross-correlation). Thus, we decided to implement, modify, extend, and integrate them in a tool ourselves.

The result is the Python library **AquDeM**[1]: **A**ctivity and Se**qu**ence **De**tection Evaluation **M**etrics, which takes two event logs in XES format as an input–one ground truth (GT) log and one log containing the detected (DET) activities–and allows for the calculation of a variety of comparison metrics for evaluation. Besides the library as core contribution, we have created a web application that utilizes the library, allowing for quick, intuitive, visual comparison of two event logs. In our research [2, 3], the Python library is not only used in this web app, but also in other more automated pipelines. The separation into library and web app allows for more varied use-cases without impacting the functionality or usability of either.

2. Innovation and Features

2.1. Library

The metrics available in AquDeM can be categorized into *activity level metrics* and *sequence level metrics*. Activity level metrics are calculated for each activity type in each case separately. A sequence level metric is calculated for each case separately, but over all activity types in that case. For calculations that span multiple cases/activities, the results are aggregated, currently using the *mean*. Another categorization is into *frame-based* or *event-based metrics* [cf. 6]. Frame-based metrics are calculated based on specific time points when an activity is detected as (in-)active, making them dependent on the (IoT) data's sampling frequency (i.e., how often data is recorded). Event-based metrics work on the classification of events themselves and do not take the sampling frequency into account. For the calculation of the metrics, the event logs must minimally adhere to these requirements: *i)* each activity execution needs both a `start` and a `complete` event; and *ii)* the logs need a sampling frequency for the frame metrics.

The metrics were selected from literature, implemented and modified based on the requirements in Section 1. To get a better understanding of the metrics, we provide intuitive (non-complete) explanations below. Note, that all of these metrics are also available in normalized form in the library, allowing for comparison among different event logs. In Table 1 we give an

[1]Video: https://youtu.be/dM4Y-80L3gA; Code: https://github.com/ics-unisg/aqudem, tags: `pkg-v0.1.1`, `fe-v0.1.1`

Table 1
Available metrics, with activity/sequence and frame/event categorization, and definition references.

Metric Abbr.	Activity/Sequence	Frame/Event	Definition
CC	Activity	Frame	cf. [7], zero-padding; input vector 1 when active, -1 when inactive
TS	Activity	Frame	cf. [6]
EA	Activity	Event	cf. [6]
LD	Sequence	Event	cf. [8]
DLD	Sequence	Event	cf. [9]

overview of the metrics regarding the categories described above, together with references to their definitions. Furthermore, we provide a usage example of AquDem in Listing 1.

- **Cross Correlation (CC)** measures the similarity between the DET and GT time series by determining the shift at which they are most alike and quantifying that similarity, relative to perfect equality for time series of that length.
- **Two Set (TS)** metrics classify frames into categories such as true positive, true negative, deletions, fragmentations, mergings, insertions, and over-fillings or under-fillings at the start and end of an activity instance.
- **Event Analysis (EA)** metrics categorize the GT events as correct, deleted, fragmented, merged, or both fragmented and merged; and DET events as correct, inserted, fragmenting, merging, or both fragmenting and merging.
- The **Levenshtein-Distance (LD)** calculates the minimum number of single activity instance edits (insertions, deletions, or substitutions) needed to transform the sequence of activity instances in DET to match GT.
- The **Damerau-Levenshtein-Distance (DLD)** extends the LD metric by also considering the transposition of two adjacent activity instances as a single edit.

```python
import aqudem
aqu_context = aqudem.Context("ground_truth_log.xes", "detected_log.xes")
aqu_context.activity_names # get all activity names present in logs
aqu_context.cross_correlation() # aggregate over all cases and activites
aqu_context.event_analysis(activity_name="Pack", case_id="1") # filter on case and activity
aqu_context.two_set(activity_name="Pack") # filter on activity, aggregate over cases
```

Listing 1: Example usage of AquDeM Python library.

2.2. Web App

The web app, built using `streamlit`,[2] has proven to be useful for the iterative and exploratory process of evaluation and development of the detection method in [2]. Notably, the library has been developed in tandem with the web app: it is built with interactive and repeated calculations

[2]https://streamlit.io/, last accessed 3rd May 2024

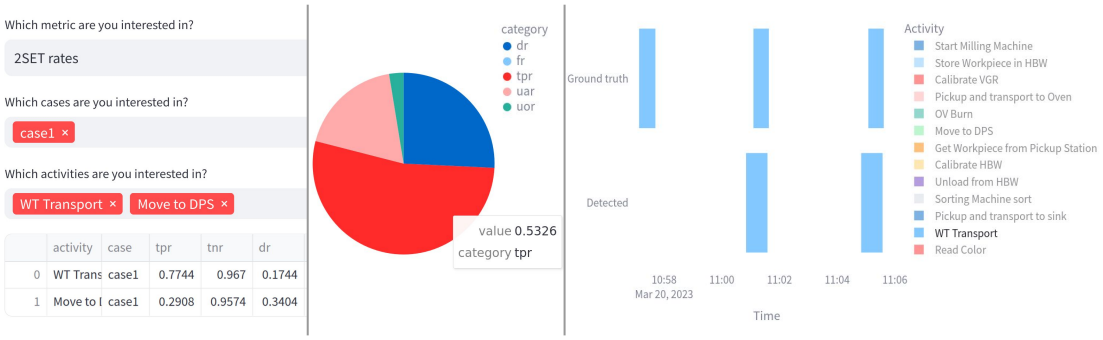

Figure 1: Screenshots of web app for event log comparison; LEFT: users can choose which metric to view, filter cases and activities; CENTER: visualizations are provided for the selected metric; RIGHT: users can view an interactive timeline of the logs, comparing detected activities with the ground truth.

in mind, i.e., browsing metrics and going back-and-forth with different analysis parameters. The library internally relies on caching to speed up recurring requests and to re-use computations from previous requests for similar requests (e.g., a filtered view that contains data calculated in a previous view). Screenshots of the web app can be seen in Figure 1. After uploading two XES logs, the user can choose a certain metric to visualize. The visualization, tabular presentation, and further options for filtering are varied for each metric to offer suitable presentations for exploration with that particular metric. The app provides specific visualizations we deemed useful (based on our use-cases), with a more flexible exploration being possible with the library.

3. Maturity and Evaluation

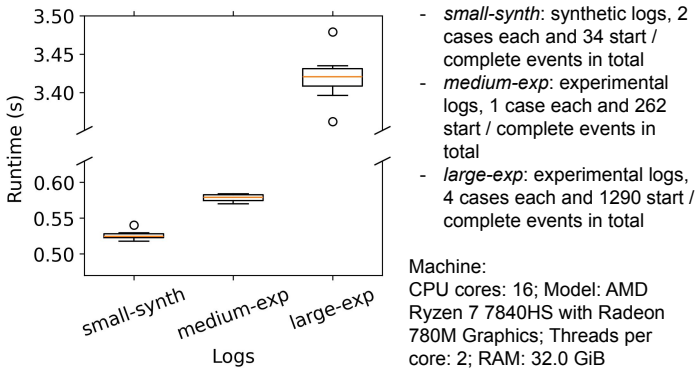

Figure 2: Boxplots (quartiles) for runtime of calculation of all available metrics over all possible case/activity combinations for ground truth/detected log pairs of varying complexity; 10 runs each.

We consider the maturity of the library and web app to be relatively high: they are continually improved and extended as they are used in our research, and include an automated test suite

with combined branch and line coverage of $> 90\%$. To better understand the library from a performance and usability perspective, we have measured runtimes with a variety of event log pairs (GT and DET), including experimental logs from the smart factory scenario described in [2, 3]. The results can be seen in Figure 2, indicating acceptable performance and scalability.

4. Conclusion

In this work we presented AquDeM: a tool featuring activity and sequence detection evaluation metrics to be used for event log comparison by BPM researchers. Besides the main, programmatically usable Python library, we provide a web app for fast, visual comparison of two event logs. The modular and decoupled design of library and web app allows for flexible usage. Given the increasing research attention in the area of activity detection in BPM and the absence of appropriate tools, we believe this to be a valuable addition to the pool of community resources.

Acknowledgments

This work has received funding from the Swiss National Science Foundation under Grant No. IZSTZ0_208497 (*ProAmbitIon* project).

References

[1] C. Janiesch, A. Koschmider, M. Mecella, B. Weber, A. Burattin, C. Di Ciccio, et al., The internet of things meets business process management: A manifesto, IEEE Systems, Man, and Cybernetics Magazine 6 (2020) 34–44.

[2] R. Seiger, M. Franceschetti, B. Weber, Data-driven generation of services for iot-based online activity detection, in: International Conference on Service-Oriented Computing, Springer, 2023, pp. 186–194.

[3] M. Franceschetti, R. Seiger, M. J. G. González, E. Garcia-Ceja, L. A. R. Flores, L. García-Bañuelos, B. Weber, Proambition: Online process conformance checking with ambiguities driven by the internet of things., in: CAiSE Research Projects Exhibition, 2023, pp. 52–59.

[4] Ieee standard for extensible event stream (xes) for achieving interoperability in event logs and event streams, IEEE Std 1849-2023 (Revision of IEEE Std 1849-2016) (2023) 1–55.

[5] F. Taymouri, M. L. Rosa, M. Dumas, F. M. Maggi, Business process variant analysis: Survey and classification, Knowledge-Based Systems 211 (2021) 106557.

[6] J. A. Ward, P. Lukowicz, H. W. Gellersen, Performance metrics for activity recognition, ACM Trans. Intell. Syst. Technol. 2 (2011).

[7] D. Lyon, The Discrete Fourier Transform, Part 6: Cross-Correlation., The Journal of Object Technology 9 (2010) 17.

[8] V. I. Levenshtein, others, Binary codes capable of correcting deletions, insertions, and reversals, in: Soviet physics doklady, volume 10, Soviet Union, 1966, pp. 707–710. Issue: 8.

[9] F. J. Damerau, A technique for computer detection and correction of spelling errors, Communications of the ACM 7 (1964) 171–176.