Scalable Predictive Process Monitoring with Multimodal Data

Predictive process monitoring (PPM) is concerned with predicting outcomes of currently running business processes. Examples for interesting outcomes of a process instance include predicting the next activity of a process instance, a binary outcome (favorable/unfavorable), or predicting the remaining runtime of a process instance. Predictions are made using machine learning models like XGBoost or neural networks (LSTMs).
During the offline phase, machine learning models are trained based on data from historical event logs which contain event data from process instances completed in the past. During the online phase, data streams from uncompleted process instances are analyzed in real-time to make predictions about the development or outcome of these process instances [1].

With the growing amount of data and the usage of additional contextual and object-centric data, existing solutions which work with event logs stored in single files (.csv) will come to their limits in the future. Therefore, new solutions which enable the scalability of PPM methods to big data will be needed. This work aims at developing a holistic approach to cover the whole pipeline of PPM, including data management, data preprocessing, and predictive model training. Previously published solutions already cover some steps of the pipeline, but these are mainly looked at in isolation.
For example, Ghahfarokhi et al. [2] presented a standard for data storage of object-centric data, which is taken up by Berti et al. [3] to implement a scalable data management solution for event logs based on MongoDB. Other work focuses on the usage of object-centric and multimodal data to improve the prediction quality of PPM methods [4, 5].

For this project, we are looking for a student who is interested in revisiting the field of PPM from a scalability perspective to detect current gaps, research and develop solutions, and implement these accordingly.

A first set of work packages as a basis for discussion is given as follows:

- Literature research about scalable data management for predictive process monitoring

- Implementation of a scalable data storage

- Development and implementation of an approach for scalable data preprocessing

- Literature Research about scalable machine learning methods

If you are interested in the topic, please send an email with your current CV and transcript of records to johannes.roider@fau.de or apply via the application form at https://www.mad.tf.fau.de/teaching/studentttheses/.

Sources:

[1] Verenich, Ilya, et al. "Survey and cross-benchmark comparison of remaining time prediction methods in business process monitoring." ACM Transactions on Intelligent Systems and technology (TIST) 10.4 (2019): 1-34.

[2] Ghahfarokhi, Anahita Farhang, et al. "OCEL: A standard for object-centric event logs." European Conference on Advances in Databases and Information Systems. Springer, Cham, 2021.

[3] Berti, Alessandro, et al. "A Scalable Database for the Storage of Object-Centric Event Logs." arXiv preprint arXiv:2202.05639 (2022).

[4] Rohrer, Timo, et al. "Predictive Object-Centric Process Monitoring." arXiv preprint arXiv:2207.10017 (2022).

[5] Cabrera, Lena, et al. "Text-Aware Predictive Process Monitoring with Contextualized Word Embeddings."