Informatik Spektrum (2022) 45:223–228
https://doi.org/10.1007/s00287-022-01470-3

HAUPTBEITRAG

Confidence-driven communication of process mining on time series

Agnes Koschmider[1] · Natascha Oppelt[2] · Marie Hundsdörfer[2]

Accepted: 15 June 2022 / Published online: 21 July 2022

Abstract
The combination of machine learning techniques with process analytics like process mining might even significantly elevate novel insights into time series data collections. To efficiently analyze time series by process mining and to convey confidence into the analysis result, requires bridging challenges. The purpose of this article is to discuss these challenges and to present initial solutions.

Introduction

In disciplines like engineering as well as life and natural sciences, time series are a common data format. Examples of time series in natural science include heights of ocean tides or seasonally occurring spikes in water temperature measures. To analyze these kinds of data, mostly correlation analysis is the first-class choice in these disciplines. Recently, the disciplines have applied machine learning techniques to advance the analysis in terms of understanding reasons behind the dataset (e.g., for a reference see [1, 2]).

The combination of machine learning techniques with process analytics like process mining might even significantly elevate novel insights into time series data collections. Process mining is a common technique concerned with automatic process analysis techniques based on event data. Such techniques include algorithms for the discovery of process models, for conformance checking between specifications and recorded events, and for predictive analytics [3]. Efficiently analyzing time series by process mining and conveying confidence into the analysis result requires bridging two challenges: (1) Virtually all techniques developed in the area of process analytics assume as input discrete data, and, at a relatively high level (i.e., close to the business level). A systematic understanding of how to efficiently bridge the gap between "raw" time series data and process analytics is needed to efficiently analyze models from time series. (2) Time series data need form and meaning to be understandable. Time series data need to be processed and represented in a useable form to turn into information and, along with this, to become an immaterial good that can lay ground for further action.

Fig. 1 shows how a potential confidence-driven communication and process analytics pipeline for time series data could look. First, the raw time series data must be pre-processed in terms of data cleaning (i.e., removing noise and outliers, extracting representative data). Then, aggregation and abstraction techniques must be applied on the pre-processed data to enhance the data with semantics, for instance, by data labeling. Next, process mining techniques could be used to discover a process by grouping or ordering the aggregated data.

Beside a process flow-based visualization, confidence measures are provided to ensure that data analysis is worthy of trust. The fusion result should be presented and visualized in an understandable (i.e., explainable) way.

In the literature, approaches exist that partially address some steps of such a confidence-driven communication and process analytics pipeline for time series data. To provide an automatic approach still requires fundamental research in the field, as will be discussed in this article. Therefore, Sect. 2 summarizes related literature referring to the analytics pipeline of Fig. 1 and indicates research challenges. Sect. 3 frames our approach within a use case. Sect. 4 positions our presented pipeline into the general idea of Cross-Domain Fusion (CDF). Sect. 5 concludes the paper.

✉ Agnes Koschmider
ak@informatik.uni-kiel.de

Natascha Oppelt
oppelt@gepgraphie.uni-kiel.de

Marie Hundsdörfer
hundsdoerfer@geographie.uni-kiel.de

[1] Group Process Analytics, Computer Science Department, Kiel University, Kiel, Germany

[2] Earth Observation and Modelling, Department of Geography, Kiel University, Kiel, Germany

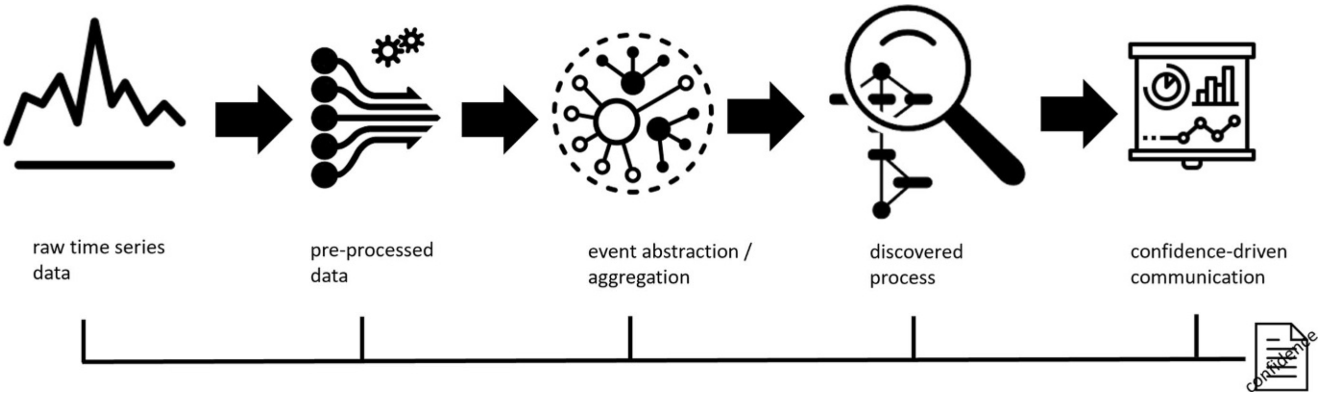

Fig. 1 Analysis pipeline from raw time series data to confidence-driven communication

Literature review

To understand the state of the art in the field, we conducted a literature review. The literature review does not claim to be exhaustive, nor has it been conducted in a systematic view. Rather, we searched for related literature surveys in the research field of interest (e.g., data pre-processing, activity discovery, confidence measures) and summarize the research challenges addressed in those surveys. The next subsections are ordered according to the topics from Fig. 1.

Challenges for raw data pre-processing

According to [4], data pre-processing consists of the following parts:

- *Data integration:* provide a consolidated view of the data from combining data from different sources
- *Data enhancement and enrichment:* enhance the data with additional information, e.g., like meta-data
- *Data transformation:* transformation of data from a source format to a required format
- *Data reduction:* reduction of the view of data like data compression
- *Data discretization:* transformation of data attributes (e.g., from continuous attributes to categorial)
- *Data cleaning:* refers to searching, identifying, and correcting data

Several literature surveys have been suggested for data pre-processing pointing to the following challenges: anomaly detection and removal [4] or efficient clustering [5].

The challenges of data pre-processing also apply to the processing of time series data. Data pre-processing is necessary before the data from the raw format can be used (i.e., time series data must be transformed into an appropriate data format), outliers must be removed, and representative data has to be extracted. Next, the processed data must be divided into subsequences representing time periods (like seasons, or weeks). Clustering is one appropriate technique to find similarities between subsequences. As a result, one subsequence can be assigned to one cluster, while one cluster is mapped onto a process activity. We call the step from a cluster to an activity an event/activity abstraction, which is summarized next.

Challenges for event/activity abstraction/aggregation

A comprehensive survey on this topic was conducted by van Zelst et al. in 2021 [3]. They classify related works in the field of event/abstraction according to four dimensions:

- *Supervision strategy* (e.g., form of supervision strategy like supervised vs. unsupervised)
- *Fine-granular event interleaving* (e.g., capability of event abstraction techniques to handle true concurrency on the higher granular level)
- *Probabilistic nature of outcome* (e.g., output of the work in terms of deterministic output)
- *Data nature* (e.g., type of data like discrete data vs. continuous data)

Based on the literature survey, the authors state that most existing techniques are supervised, the notion of continuous data sources with the notion of process instances is hardly tackled in the literature, the majority of the techniques results in a discrete output, and most techniques work on offline data rather than online. This gives room for future techniques relying on unsupervised learning, addressing event interleaving and area of probabilistic output models, and works for online data settings. Next, the activities are taken as input for process discovery or process mining (see step "discovered process" in Fig. 1).

Table 1 Summary of challenges related to confidence-driven communication of process discovery on time series data

Data pre-processing	Event/abstraction aggregation	Process discovery on raw data	Confidence-driven communication
Efficient pre-processing in terms of data integration, data enhancement and enrichment, data transformation, data reduction, data discretization, and data cleaning	Future works should support techniques relying on unsupervised learning, address event interleaving and area of probabilistic output models, and work for online data settings	Define and adjust the spectrum of methods that can discover process flows from time series data	Provide a reflection on the impact of these explanation methods to provide confidence and trust for visualizations for time series

Challenges for process mining on aggregated (raw) data

The discovery of process models on data or events that have been abstracted from raw data like sensor event data as well as time series or video data is currently highly requested. As discussed in the previous section, several techniques have been suggested to tackle event abstraction. The abstracted events then need to be aggregated to an event log, and process mining is applied. The BPM-IoT manifesto [6] discusses several challenges concerned with process mining on higher levels of knowledge (i.e., in our context abstracted events) (see Fig. 1 in [6]):

- *Integrating raw data into the correctness check of processes:* the discovered processes should specifically consider the raw nature of some components.
- *Detecting new processes from data:* how to consider situational knowledge (specific for raw data) into process discovery.
- *Dealing with new situations:* how to consider ad-hoc (real-time) decisions into process discovery implicating a continuous change of processes.
- *Specifying the autonomy of new things:* how to integrate the concept of autonomy (i.e., to grant things full autonomy to decide) into process discovery.

Related works on online process mining [7, 8] partially address these challenges. However, these approaches do not consider time series data. Thus, research is required to understand how to transfer existing knowledge and to efficiently bridge the gap between raw time series data and process mining.

Challenges for confidence-driven communication

Generally, confidence measures the risk as to how sure users are that they received the correct suggestions by the (artificial intelligence-based) model. Confidence measures have not gathered much attention as yet, although they are a key ingredient to building trustworthy systems. Particularly in domains where automated data analytics techniques are not yet common, corresponding methods could then ensure that data analysis is worthy of trust. To ensure that machine learning models in the natural sciences are trustworthy, confidence should be prioritized. Our literature analysis indicates research needs reflecting impacts between confidence, trust and analysis results. Here, data storytelling and explainability measures might be an appropriate solution, and these are worth investigating in the future (Table 1).

Use case

To develop a (semi)automatic pipeline for confidence-driven time series analysis for process mining, we aim to use data from the bacteria genus *Vibrio* as a use case. These species are ubiquitous members of coastal, estuarine, and brackish environments and become a threat for swimmers. The most popular kind will be *V. cholerae O1/O139* causing severe cholera disease. Increased water temperatures lead to increased population densities of *V. vulnificus* and the occurrence of potentially pathogenic strains of other *non-cholerae Vibrios*, which may cause infections in immunodeficient humans. Recent studies indicate that water temperature and salinity govern the occurrence (and the massive increase) of *non-cholerae Vibrios* [9]. Although water temperature and salinity seem to play a major role for the occurrence of *Vibrios*, other factors such as chlorophyll and the occurrence of aquatic vegetation [10, 11] associated with a higher carbon content [12] seem to influence the frequency of massive blooms.

Most infections occur in subtropical regions between April and November [13]. In the context of global warming and an increase in heat waves, however, infections have increased and will probably further increase due to the changing water conditions. An increase in severe infections and fatalities are also seen in regions such as the Baltic Sea area [14].

In Germany, reporting obligations of infections were absent until recently. Therefore, systematic recordings about infections and fatalities due to *non-cholerae Vibrios* were lacking until 2019. Most infections occurred during the hot summers of 2014 and 2015. In the two years 2018 and 2019, there were exceptional summer heat waves, during which a total of 81 cases of infection were published by [15]. The low salinity and rapidly warming body of water of the Baltic Sea provides an ideal breeding ground for

Table 2 Pre-processed data for the use case of the bacteria genus *Vibrio*

Date	Monitoring site index	Monitoring site	Region	Coordinates	*V. para.* (*n*)	*V. vuln.* (*n*)	*V. chol.* (*n*)	Total *Vibrio*	Air temperature (°C)	Sea surface temp (°C)	Sea surface salinity (psu)
2014-05-06	MS_46	SH Brokdorf Elbe Sandstrand bathing site ID DESH_PR_0319	Elbe	(53.8591, 9.3287)	0	16	0	16	17,06	17,11	29,58
2015-08-10	MS_49	SH Schlei Kappeln	Schlei	(54.6765, 10.0353)	38	1	0	39	18,59	18,27	32,30
2015-08-24	MS_51	Schlei Borgwedel Jugendherberge DESH_PR_0268	Schlei	(54.49761691, 9.6722786)	0	39	23	62	18,75	19,06	30,28
2015-08-31	MS_51	Schlei Borgwedel Jugendherberge DESH_PR_0268	Schlei	(54.49761691, 9.6722786)	3	20	53	76	17,55	18,02	33,90
2015-09-07	MS_53	Schlei Selker Noor Niederselk DESH_PR_0270	Schlei	(54.4790552, 9.582782934)	0	0	666	666	12,89	16,84	34,10
2016-07-25	MS_60	SH Eider Wollersum	Eider	(54.33124465811813, 8992625376823417)	1	17	8	26	20,53	18,76	29,08
2016-08-08	MS_58	SH Eider Eidersperrwerk	Eider	(54.26474192264182, 8.845830796878813)	3	0	0	3	16,70	15,16	35,43
2018-06-11	MS_65	SH Gammendorf bathing site ID 0115	Fehmarn	(54.52805978, 11.11392336)	5	0	10	15	16,70	17,38	19,45
2018-08-06	MS_64	SH Fehmarnsund bathing site ID 0117	Fehmarn	(54.40187556, 11.15274118)	25	82	0	107	20,59	22,64	36,11
2019-08-19	MS_21	SH OSTS GROSSENBRODE KURZENTRUM bathing site ID DESH_PR_0095	Fehmarn	(54.35737828, 11.08880255)	1	1	0	2	17,53	17,69	32,24

V. vulnificus, which is mainly responsible for infections and deaths of patients [16].

To sum up, the data source for the use case is as follows: we have sparse spatio-temporal data and context information related to the data. Table 2 shows an excerpt of the data set to be used as input for our analytics pipeline. The data has already been cleaned in terms of completing missing entries and unifying the data. To identify the correlations and causalities between parameters affecting the occurrence of the *Vibrio* bacteria, we aim to apply the analysis pipeline as shown in Fig. 1 and to provide a decision-making system ensuring confidence. The benefits of process mining for this use case is to identify causalities between variables (i.e., does temperature have an impact on infection risk?). In this way, our pipeline makes it possible to understand a temporal pattern/trend in what is being measured and to identify outliers that can help prevent unintended consequences and point to new processes.

Connection to the main idea of cross-domain fusion

Key elements for a distinct analysis and forecast of *V. vulnificus* abundance is the fusion and use of data from different sources and disciplines. Since the combination of several factors seems to influence the abundance of these bacteria, we need to investigate the impact of the above-mentioned parameters. Besides data fusion, time series are a key data format used in disciplines like engineering as well as life and natural sciences. Also, data sampling is necessary to efficiently support event/activity abstractions (e.g., like near aquatic vegetation for seagrass and Sargassum).

Conclusion and outlook

The article summarized research challenges for a confidence-driven analysis of "raw" time-series data for process mining. To provide a solution we plan to re-evaluate the spectrum of data pre-processing techniques that can be utilized to efficiently process time series for data and process fusions, define techniques of event/activity abstraction, enhance process mining techniques to discover a process structure, and build an understanding of how users (i.e., researchers, students, laymen) of time series visualizations engage with that information, learn more about scientific work practices, and how they interpret results. Particularly, we plan to investigate these issues for the occurrence of the bacteria genus *V. vulnificus* in coastal waters of the Baltic Sea to develop, observe, and visualize risk of infections useful for local stakeholders and health public authorities.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Wei X, Yang H-Q, Zhang L, Yao Y-P (2020) Machine learning for pore-water pressure time-series prediction: application of recurrent neural networks. Geosci Front. https://doi.org/10.1016/j.gsf.2020.04.011
2. Torrisi M, Gianluca Pollastri Le Q (2020) Deep learning methods in protein structure prediction. Comput Struct Biotechnol J. https://doi.org/10.1016/j.csbj.2019.12.011
3. van Zelst SJ, Mannhardt F, de Leoni M et al (2021) Event abstraction in process mining: literature review and taxonomy. Granul Comput 6:719–736. https://doi.org/10.1007/s41066-020-00226-2
4. Zhang A, Song S, Wang J, Yu PS (2017) Time series data cleaning: from anomaly detection to anomaly repairing. Proc Vldb Endow 10(7):1046–1057. https://doi.org/10.14778/3115404.3115410
5. Esling P, Agon C (2012) Time-series data mining. ACM Comput Surv. https://doi.org/10.1145/2379776.2379788
6. Janiesch, Koschmider A, Mecella M, Weber B, Burattin A, Di Ciccio C et al (2020) The Internet of things meets business process management: a manifesto C IEEE systems. Man Cybern Mag 6(4):34–44
7. Schuster D, van Zelst SJ (2020) Online process monitoring using incremental state-space expansion: an exact algorithm. BPM 12168:147–164
8. Burattin A, Cimitile M, Maggi FM, Sperduti A (2015) Online Discovery of Declarative Process Models from Event Streams. IEEE Trans Serv Comput 8(6):833–846
9. Baker-Austin C, Oliver JD (2018) Vibrio vulnificus. New insights into a deadly opportunistic pathogen. Environ Microbiol 20(2):423–430. https://doi.org/10.1111/1462-2920.13955
10. Reusch TBH, Schubert PR, Marten S-M, Gill D, Karez R, Busch K, Hentschel U (2021) Lower Vibrio spp. abundances in Zostera marina leaf canopies suggest a novel ecosystem function for temperate seagrass beds. Mar Biol 168:149. https://doi.org/10.1101/2021.03.21.436319
11. Michotey V, Blanfuné A, Chevalier C, Garel G, Diaz F, Berline L, Le Grand L, Armougom F, Guasco S, Ruitton S, Changeux T, Belloni B, Blanchot J, Ménard F, Thibaut T (2020) In situ observations and modelling revealed environmental factors favouring occurrence of Vibrio in microbiome of the pelagic Sargassum responsible for strandings. Sci Total Environ 748(2020):1216
12. Oberbeckmann S, Fuchs BM, Meiners M et al (2012) Seasonal dynamics and modeling of a Vibrio community in coastal waters of the North Sea. Microb Ecol 63(3):543–551. https://doi.org/10.1007/s00248-011-9990-9
13. Li M, Zhao L, Ma J et al (2018) Vibrio vulnificus in aquariums is a novel threat to marine mammals and public health. Transbound Emerg Dis 65(6):1863–1871. https://doi.org/10.1111/tbed.12967

14. Baker-Austin C, Trinanes J, Gonzalez-Escalona N et al (2017) Non-cholera Vibrios. The microbial barometer of climate change. Trends Microbiol 25(1):76–84. https://doi.org/10.1016/j.tim.2016.09.008

15. Brehm TT, Dupke S, Hauk G et al (2021) Nicht-Cholera-Vibrionen – derzeit noch seltene, aber wachsende Infektionsgefahr in Nord- und Ostsee. Internist 62:876–886. https://doi.org/10.1007/s00108-021-01086-x

16. Metelmann C, Metelmann B, Gründling M et al (2020) Vibrio vulnificus, eine zunehmende Sepsisgefahr in Deutschland? Anaesthesist 69:672–678. https://doi.org/10.1007/s00101-020-00811-9