Temporal Deviations on Event Sequences

Janina Sontheim, Florian Richter, and Thomas Seidl

Ludwig-Maximilians-Universität, Munich, Germany
{sontheim, richter, seidl@dbs.ifi.lmu.de}

Abstract. Time deviations in business processes - depending on gradient and severity - are crucial for the performance of a business and finally leading to a gain or a loss in return and reputation. Therefore, focusing on the time perspective of processes in general is very important. Even more important is the temporal behavior of a single execution, a case, to find difficulties and potential in the process. The identification of cases that differ from the default execution allows to understand individual instances.
Conservative data mining on event sequences, so called process mining, exclusively focuses on structural aspects of the process. These approaches, however, are unaware of temporal aspects regarding accelerations or decelerations of activity execution times and neglect a very powerful adjusting screw. Our novel signature for cases tackles this task by representing cases depending on their temporal deviation behavior. Thus processes with their cases can be monitored on a entirely new level and anomalies and derivations regarding time can be identified.

Keywords: Process Mining · Case Profiling · Time Deviation.

1 Introduction

Determining the conformance of a singular case is one of the three key tasks in process mining beside process discovery and model enhancement. It confirms case compliance to the assumed underlying process model structure. Agrawal et al. were the first who tackled this topic in process mining in [2]. Senderovich et al. dealed in [6][5] with mining of process delays on event-level. They focused on structural deviations from the baseline process which is also the issue with all current conformance checking approaches and neglect thereby the temporal perspective. We focus on the temporal perspective and thus improve the spectrum of mining possibilities in the field of conformance checking.

For example considering a manufacturing process with some artisans producing chairs, as skeched in Fig. 1. Every artisan *crafts components* first. Then they *assemble* all the pieces to build the whole *chair*. Finally, the chair is checked and *deficiencies get remedied*. In the example in Fig. 1 there are two prominent variants which do temporally not conform to the preceding cycles. Here the so far published conformance checking approaches can only detect that both cases act conform to the desired process.

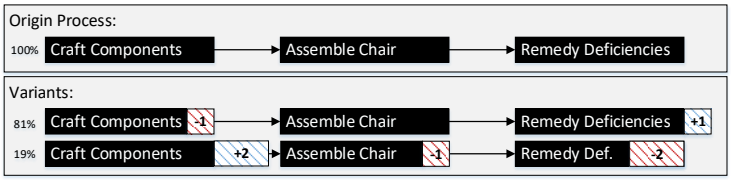

Fig. 1. Partitioning the process into two variants reveal, that a minority of cases needs less time and should be selected as a best practice for future cases.

Another process mining key task is model enhancement. Of our special interest is the direction regarding the temporal perspective, for which Cheikhrouhou et al. wrote a survey [3] outlining existing approaches and ongoing research challenges. A subarea of it is time prediction. Van der Aalst et al. presented in [1] an approach to predict the completion time of running instances. Though time predictions are currently just for single events ignoring the existence of time correlation between events of the same case. In our previous example e.g. there is a dependency between the event *Crafts Components* and *Remedy Deficiencies*. With our temporal deviation signatures this field gets a new perspective due to considering event dependencies.

One of our strongest areas of application is clustering which is a further topic in process mining. Related to clustering is the topic of anomaly detection. Rogge-Solti and Kasneci covered in [4] an anomaly detection approach with a temporal perspective, though it is done after the process model is created from the log whereby it can just find temporal anomalies regarding the process model. Our approach uses the log file to calculate the signature and hence it does not contain any process modeling errors. Our signature can support clustering with an additional view. Each case might be different but there are other cases that are similar to the viewed case. Using our new signature for cases we are able to define a clustering regarding time aspects and thus among others remaining time prediction will become more precise.

Initially we focus on clustering of cases regarding temporal aspects. Therefore we urgently need a signature for cases to be able to cluster them. While representations of structural deviations of cases are researched in process mining, a representation of temporal signatures for cases was not investigated to the best of our knowledge.

2 Case Signatures for Temporal Deviations

To formally introduce events and logs we start with defining the activity space $\mathcal{A}$ as the set of actions which can occur during a process execution. An *event* $e = (c, a, t)$ is then defined as an aggregation of a *case identifier* $c \in \mathbb{N}$, an *activity* $a \in \mathcal{A}$ and a *timestamp* $t \in \mathbb{N}$. The event space is denoted as $\mathcal{E}$. The set of all events containing the same case identifier is called *case*. Some abbreviations will help to keep the following explanations clearer: For any event $e = (c, a, t)$ we

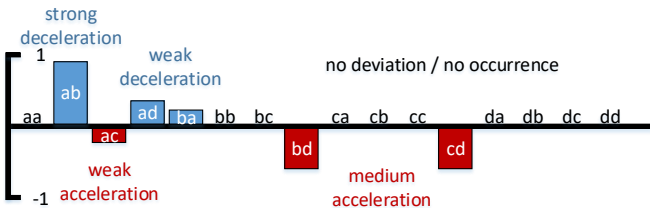

Fig. 2. The graphical representation of a temporal deviation signature. The signature contains all possible pairs for four activities a, b, c, d, plotted on the x-axis. The y-axis corresponds to the z-scoring.

define $e.c = c$, $e.a = a$ and $e.t = t$. A log L is a (multi-)set of events resp. a set of cases.

We compute time intervals between pairs of activities within the same case. Only considering directly consecutive activities should be avoided due to potential temporal correlations. Instead we consider all succeeding activity pairs within a case. This step requires quadratic effort. If the application consists of rather long cases, it is reasonable to trade performance for accuracy. Each trace can now be mapped to a vector of durations, as shown in Fig. 2, where blue bars on top of the zero line indicate longer durations and red bars below indicate shorter durations relative to the process means. The dimensions of the vectors differ depending on the case lengths.

As the aim of the signature is among others to support finding clusters of traces it is very convenient that any clustering method for vector data can be applied at this state. Since we are especially interested in abnormal temporal behavior we use a normalization focusing on deviations. Thus we use z-scoring as it puts emphasis on the degree of deviation by normalizing with an attributes' standard deviation and mean value. Although this is mostly useful for Gaussian distributed values, it works quite fine in process data due to the large amount of events in most process logs. This allows us to utilize the central limit theorem of statistics if we assume that activities among different cases are mostly independently and identically distributed. Working with values relative to their variance instead of absolute values counters vastly the balancing of dimensions.

In many applications the gradient of small deviations is more important than large differences between large values. A small deviation of few minutes can already point towards a major problem while it does rarely matter if an event is delayed by 12 hours or by 24 hours. Here we use again a simple method by applying a sigmoid function. The particular choice is not very important so we apply the fastest one: $S(x) = x/(1 + |x|)$. Applying these steps leads us to a vector representation of a case containing temporal and structural properties. One should keep in mind that the common vector space of all process instances has a very high dimension although each case exists only in a lower dimensional subspace.

Case	Activity	Timestamp
1	A	0
2	A	0
3	A	0
4	A	0
5	A	0
3	B	2.8
1	B	3.3
5	B	4.4
2	B	4.5
4	B	5.0
3	C	7.3
1	C	8.1
2	C	8.9
4	C	9.2
5	C	10.0

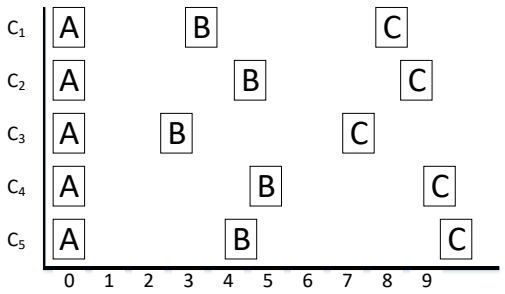

Fig. 3. An example process log contains 5 short traces with the same structure but different temporal behavior, although the start time is always the same. This log is sorted by the timestamp.

Given a process with activity space $\mathcal{A}$ we call $\mathcal{R} = [-1, 1]^{|\mathcal{A}\times\mathcal{A}|}$ the deviation signature space. Let μ be the mean, σ the standard deviation of the chosen distribution, and $(c, a_i, t_i), (c, a_j, t_j) \in C$. Then we define the **temporal deviation signature of case c**, with $c = \langle (c, a_1, t_1), \ldots, (c, a_n, t_n) \rangle$, as the vector $\boldsymbol{v_c} \in \mathcal{R}$:

$$v_c(a_i, a_j) = \begin{cases} S\left(\frac{|t_i - t_j| - \mu_{(a_i,a_j)}}{\sigma_{(a_i,a_j)}}\right) & , i < j \\ 0 & , \text{otherwise.} \end{cases}$$

To illustrate the previous steps in an example we give a small sample log in Figure 3 with a schematic representation. The mean value and the standard deviation of all activity pairs is then computed: $\mu_{(a,b)} = 4.0$, $\sigma_{(a,b)} = 0.82$, $\mu_{(a,c)} = 8.7$, $\sigma_{(a,c)} = 0.93$, $\mu_{(b,c)} = 4.7$, $\sigma_{(a,b)} = 0.49$. For case 3 we compute the temporal deviation signature consisting of the activity pairs (ab, ac, bc). The relevant interim times are $(2.8, 7.3, 4.5)$ and the z-scoring is $(-1.47, -1.51, -0.41)$. After the application of S we receive the final temporal deviation signature $(-0.59, -0.60, -0.29)$. For all cases 1 to 5 of this example process the temporal deviation signatures are shown in comparison ordered as (ab, ac, bc):

$$v_1 = \begin{pmatrix} -0.46 \\ -0.39 \\ 0.17 \end{pmatrix} \quad v_2 = \begin{pmatrix} 0.38 \\ 0.18 \\ -0.38 \end{pmatrix} \quad v_3 = \begin{pmatrix} -0.59 \\ -0.60 \\ -0.29 \end{pmatrix} \quad v_4 = \begin{pmatrix} 0.55 \\ 0.35 \\ -0.51 \end{pmatrix} \quad v_5 = \begin{pmatrix} 0.33 \\ 0.58 \\ 0.65 \end{pmatrix}$$

3 Research Directions

With the temporal deviation signature we present a novel process representation which puts emphasis on the very important temporal view. The identification of temporal deviations reveals great insights and improves a process or avoids drawbacks. For future directions a clustering of cases with similar temporal deviation signatures can be modeled to reveal various clusters of process variants across a whole process which leads to great possibilities for businesses.

References

1. Van der Aalst, W.M., Schonenberg, M.H., Song, M.: Time prediction based on process mining. Information Systems **36**(2), 450–475 (2011)
2. Agrawal, R., Gunopulos, D., Leymann, F.: Mining process models from workflow logs. In: EDBT. pp. 469–483 (1998)
3. Cheikhrouhou, S., Kallel, S., Guermouche, N., Jmaiel, M.: The temporal perspective in business process modeling: a survey and research challenges. Service Oriented Computing and Applications **9**(1), 75–85 (2015)
4. Rogge-Solti, A., Kasneci, G.: Temporal anomaly detection in business processes. In: International Conference on Business Process Management. pp. 234–249. Springer (2014)
5. Senderovich, A., Weidlich, M., Gal, A.: Temporal network representation of event logs for improved performance modelling in business processes. In: International Conference on Business Process Management. pp. 3–21. Springer (2017)
6. Senderovich, A., Weidlich, M., Yedidsion, L., Gal, A., Mandelbaum, A., Kadish, S., Bunnell, C.A.: Conformance checking and performance improvement in scheduled processes: A queueing-network perspective. Information Systems **62**, 185–206 (2016)