Information Systems 46 (2014) 102–122

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

Single-Entry Single-Exit decomposed conformance checking

Jorge Munoz-Gama [a,*], Josep Carmona [a], Wil M.P. van der Aalst [b,c]

[a] Universitat Politecnica de Catalunya, Barcelona, Spain
[b] Eindhoven University of Technology, Eindhoven, The Netherlands
[c] PAIS Lab, Higher School of Economics, Moscow, Russia

ARTICLE INFO

Available online 1 May 2014

Keywords:
Process mining
Conformance checking
Decomposition
Process diagnosis

ABSTRACT

An exponential growth of event data can be witnessed across all industries. Devices connected to the internet (internet of things), social interaction, mobile computing, and cloud computing provide new sources of event data and this trend will continue. The omnipresence of large amounts of event data is an important enabler for process mining. Process mining techniques can be used to discover, monitor and improve real processes by extracting knowledge from observed behavior. However, unprecedented volumes of event data also provide new challenges and often state-of-the-art process mining techniques cannot cope. This paper focuses on “conformance checking in the large” and presents a novel decomposition technique that partitions larger process models and event logs into smaller parts that can be analyzed independently. The so-called Single-Entry Single-Exit (SESE) decomposition not only helps to speed up conformance checking, but also provides improved diagnostics. The analyst can zoom in on the problematic parts of the process. Importantly, the conditions under which the conformance of the whole can be assessed by verifying the conformance of the SESE parts are described, which enables the decomposition and distribution of large conformance checking problems. All the techniques have been implemented in ProM, and experimental results are provided.

1. Introduction

In the last decade process mining emerged as a novel discipline for addressing challenges related to Business Process Management (BPM) and “Big Data” [1]. Information systems (and many other computer-supported systems) record overwhelming amounts of event data. These can be seen as the “footprints” left by the process. For example, Boeing jet engines may produce up to 10 terabytes of operational information every 30 min, and Walmart logs may store 1 million customer transactions per hour [2].

Event logs can be used to conduct three types of process mining [1]. The first and most prominent is *discovery*. A discovery technique takes an event log and produces a model without using a priori information. For many organizations it is surprising that existing techniques are indeed able to discover real processes based only on example behaviors recorded in event logs. The second type is *conformance* where an existing process model is compared with an event log of the same process. Conformance checking can be used to check if reality, as recorded in the log, conforms to the model and vice versa. The third type is *enhancement* where the idea is to extend or improve an existing process model using information about the actual process recorded in an event log. Whereas conformance checking measures alignment between model and reality, this third type of process mining aims

* Corresponding author. Tel.: +34 93 4137861; fax: +34 93 4137833.
E-mail addresses: jmunoz@lsi.upc.edu (J. Munoz-Gama), jcarmona@lsi.upc.edu (J. Carmona), w.m.p.v.d.aalst@tue.nl (W.M.P. van der Aalst).

http://dx.doi.org/10.1016/j.is.2014.04.003
0306-4379/© 2014 Elsevier Ltd. All rights reserved.

to change or extend the a priori model; for instance, using timestamps in the event log, one can extend the model to show bottlenecks, service levels, throughput times, and frequencies.

In conformance checking, the seminal work by Rozinat et al. [3] was the first in formalizing the problem and enumerating the four dimensions to consider for determining the adequacy of a model in describing a log: *fitness*, *precision*, *generalization* and *simplicity*. In this paper, we will focus on evaluating fitness that measures the capability of a model in reproducing the traces of a log. As modeling notation, we will focus on the Petri net formalism [4], although most of the conclusions of the paper can be generalized to similar process formalisms [5].

In real-life situations, event logs often do not fit its corresponding models, i.e., some log traces cannot be fully reproduced in the model. These non-fitting situations should be communicated to the stakeholders, in order to take decisions on the process object of study. However, in reality process models can be non-deterministic, which complicates the analysis. Non-determinism may arise when the model contains *silent* or *duplicate* activities, which is often the case in practice. Moreover, the presence of *noise* in the log (rare or infrequent behavior that has been recorded in the log) complicates even more the algorithmic detection of non-fitting situations. Due to this, the initial proposal from Rozinat et al. to *replay* log traces in a model in order to assess whether a trace can fit a model has been recently reconsidered, giving rise to the notion of *alignment*. Alignment techniques relate execution sequences of the model and traces in the event log. The techniques can cope with deviations and models with duplicate/invisible activities [6–8]. However, alignment techniques are extremely challenging from a computational point of view. Traces in the event log need to be mapped on paths in the model. A model may have infinitely many paths and the traces may have an arbitrary amount of deviating events. Hence, although the algorithms have demonstrated to be of great value for undertaking small- or medium-sized problem instances [1,9], they are often unable to handle problems of industrial size. We believe that decomposition techniques are an important means to tackle much large and more complex process mining problems. Therefore, this paper addresses this problem through *decomposition and distribution*.

There is a trivial way to decompose the conformance checking problem. One can simply split the event log into sublogs such that every trace appears in precisely one of these sublogs. Note that the conformance is still checked on the whole model. Linear speed-ups are possible using such a simple decomposition. However, the real complexity is in the size of the model and the number of different activities in the event log. Therefore, we propose a different approach. Instead of trying to assess the conformance of the whole event log and the complete Petri net, conformance checking is only performed for selected subprocesses (subnets of the initial Petri net and corresponding sublogs). Subprocesses are identified as subnets of the Petri net that have a Single-Entry and a Single-Exit (*SESE*) node, thus representing an isolated part of the model with a well-defined interface to the rest of the net. SESEs can be efficiently computed and hierarchically represented in a tree-like manner into the Refined Process Structured Tree (RPST) [10].

Experiments (cf. Section 6) show a *considerable reduction (orders of magnitude) in the time required to perform fitness checking*. Moreover the techniques presented in this paper allow for identifying those subnets that have fitness problems, allowing the process owner to focus on the problematic parts of a large model. Importantly, we have performed analytical comparisons of the conformance problem when decomposition is considered or not, related to the size of the components and the average length of the log traces. In terms of complexity, those studies reveal a clear superiority of the methods proposed in this paper, being more robust for these important matters (size of the components and length of log traces). Remarkably, this significant complexity alleviation comes without any penalty on the capability of the method: by applying decomposition techniques conformance checking of the whole can still be assessed.

The SESE decomposition is not only used for efficiency reasons. We also use it to provide *diagnostics* that help the analyst in localizing conformance problems. We create a topological structure of SESEs in order to detect the larger connected components that have fitness problems. Moreover, problematic parts can be analyzed in isolation. Finally, a hierarchical perspective of the conformance checking problem is presented, which may open the door for zoom-in zoom-out analysis, and also focus the analysis of the hierarchy into particular subprocesses that have common features.

This paper extends and generalizes two recent conference papers [11,12]. The extensions and generalizations can be summarized as follows. First, we present a strategy to compute fitness by adapting a partitioning of the RPST in order to satisfy the valid decomposition requirements from [13]. Second, we have extended the decomposition approach of [12] in order to deal with silent and duplicate activities. Third, different perspectives to the conformance problem are presented which aim to provide more flexibility during analysis. Three alternatives to the traditional conformance checking practice are proposed: (1) subprocess, (2) hierarchical and (3) filtered conformance checking. The filtered conformance checking perspective is based on the initial one presented in [11], but a new data perspective is proposed in this paper. Fourth, we have reimplemented the initial architecture of [11,12] in order to address problems encountered when analyzing large event logs. The new implementation results in some cases in speed-ups of orders of magnitude. Fifth, we have extended considerably the empirical evaluation from [11,12], incorporating studies that relate the performance of the decomposed technique with respect to the log trace length and the size of the subprocesses.

The paper is structured as follows: Section 2 introduces preliminaries needed in the remainder. In Section 3 the SESE-decomposition is presented formally. Section 4 describes the high-level use of the RPST structure for diagnostics of conformance checking problems, by means of a topology of SESE components. Various applications of the decomposition technique are presented in Section 5.