Information Systems 56 (2016) 235–257

Contents lists available at ScienceDirect

Information Systems

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

A general process mining framework for correlating, predicting and clustering dynamic behavior based on event logs

Massimiliano de Leoni [a,*], Wil M.P. van der Aalst [a], Marcus Dees [b]

[a] *Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands*
[b] *Uitvoeringsinstituut Werknemersverzekeringen (UWV), The Netherlands*

ARTICLE INFO

Available online 17 July 2015

Keywords:
Process mining
Decision and regression trees
Event-log manipulation
Event-log clustering

ABSTRACT

Process mining can be viewed as the missing link between model-based process analysis and data-oriented analysis techniques. Lion's share of process mining research has been focusing on process discovery (creating process models from raw data) and replay techniques to check conformance and analyze bottlenecks. These techniques have helped organizations to address compliance and performance problems. However, for a more refined analysis, it is essential to *correlate different process characteristics*. For example, do deviations from the normative process cause additional delays and costs? Are rejected cases handled differently in the initial phases of the process? What is the influence of a doctor's experience on treatment process? These and other questions may involve process characteristics related to different perspectives (control-flow, data-flow, time, organization, cost, compliance, etc.). Specific questions (e.g., predicting the remaining processing time) have been investigated before, but a generic approach was missing thus far. The proposed framework unifies a number of approaches for correlation analysis proposed in literature, proposing a general solution that can perform those analyses and many more. The approach has been implemented in ProM and combines process and data mining techniques. In this paper, we also demonstrate the applicability using a case study conducted with the UWV (Employee Insurance Agency), one of the largest "administrative factories" in The Netherlands.

1. Introduction

Process Aware Information Systems (PAISs) are increasingly used by organizations to support their businesses. Some of these systems are driven by process models, e.g., Business Process Management (BPM) and Workflow Management (WFM) systems. However, in most PAISs only an *implicit* process notion exists. Consider for example the enterprise software SAP and Oracle that is widely used to manage business operations and customer relations. Although these systems provide BPM/WFM functionality, most processes are partly hard-coded in application software and exist only in the minds of the people using the software. This provides flexibility and simplifies implementation efforts, but also results in poor management support. If there is no *explicit* process model that reflects reality, it is impossible to reason about compliance and performance in a precise and unified manner. Management dashboards provided by Business Intelligence (BI) software tend to oversimplify reality and do not use explicit process models. Fortunately, all these systems record the execution of process instances in so-called *event logs*. These logs thus capture information about

[*] Corresponding author.
E-mail addresses: m.d.leoni@tue.nl (M. de Leoni), w.m.p.v.d.aalst@tue.nl (W.M.P. van der Aalst), marcus.dees@uwv.nl (M. Dees).

http://dx.doi.org/10.1016/j.is.2015.07.003
0306-4379/© 2015 Elsevier Ltd. All rights reserved.

activities performed. Each event records the execution of an activity instance by a given resource at a certain point in time along with the output produced.

Event logs are the key enabler for *process mining*, which is capable of extracting, from event logs, in-depth insights in process-related problems that contemporary enterprises face [1]. Through the application of process mining, organizations can discover the processes as they were conducted in reality, check whether certain practices and regulations were really followed and gain insight into bottlenecks, resource utilization, and other performance-related aspects of processes.

Process mining often starts with *process discovery*, i.e., automatically learning process models based on raw event data. Once there is a process model (discovered or made by hand), the events can be replayed on the model to *check conformance* and to *uncover bottlenecks* in the process [1]. However, such analyses are often only the starting point for providing initial insights. When discovering a bottleneck or frequent deviation, one would like to understand why it exists. This requires the correlation of different *process characteristics*. These characteristics can be based on:

- the *control-flow* perspective (e.g., the next activity going to be performed);
- the *data-flow* perspective (e.g., the age of the patient or the amount of glucose in a blood sample);
- the *time* perspective (e.g., the activity duration or the remaining time to the end of the process);
- the *resource/organization* perspective (e.g., the resource going to perform a particular activity or the current workload), or,
- if a normative process model exists, the *conformance* perspective (e.g., the skipping of a mandatory activity or executing two activities in the wrong order).

There are of course other perspectives possible (e.g., the cost perspective), but these are often not process-specific and can be easily encoded in the data-flow perspective. For example, there may be data attributes capturing variable and fixed costs of an activity.

These problems are specific instances of a more general problem, which is concerned with *relating any process or event characteristic to other characteristics associated with single events or the entire process*. This paper proposes a framework to solve the more general correlation problem and provides a very powerful tool that unifies the numerous ad hoc approaches described in literature. This is achieved by providing (1) a broad and extendable set of characteristics related to control-flow, data-flow, time, resources, organizations and conformance, and (2) a generic framework where any characteristic (dependent variable) can be explained in terms of correlations with any set of other characteristics (independent variables). For instance, the involvement of a particular resource or routing decision can be related to the elapsed time, but also the other way around: the elapsed time can be related to resource behavior or routing.

The approach is fully supported by a new package that has been added to the open-source process mining framework *ProM*.[1] The evaluation of our approach is based on two case studies involving UWV, a Dutch governmental institute in charge of supplying benefits. For the first case study, the framework allows us to successfully answer process-related questions related to causes of observed problems within UWV (e.g., reclamations of customers). For some problems, we could show surprising root causes. For other problems, we could only show that some suspected correlations were not present, thus providing novel insights. A second case study was concerned with discovering the process model that describes the UWV's management of provisions of benefits for citizens who are unemployed and unable to look for a job for a relatively long period of time because of physical or mental illnesses. Analysis shows that there is a lot of variability in process execution, which is strongly related to the length of the benefit's provision. Therefore, the discovery of one single process led to unsatisfactory results as this variability cannot be captured in a single process model. After splitting the event log into clusters based on the distinguishing features discovered through our approach, the results significantly improved.

It is important to note that our framework does *not* enable analyses that previously were not possible. The novelty of our framework is concerned with providing a single environment where a broad range of process-centric analyses can be performed much quicker without requiring process analysts with a solid technical background. Without the framework, for instance, process analysts are confronted with database systems where they need to perform tedious operations to import data from external sources and, later, to carefully design complex SQL queries to perform joins and even self-joins that involve different database tables and views.

The remainder of the paper is as follows. Section 2 presents the overall framework proposed in this paper. The section also discusses how the problem of relating business process characteristics can be formulated and solved as a data-mining problem. In particular, we leverage of data-mining methods to construct a decision or regression tree. In the remainder, we refer them to as prediction trees if there is no reason to make a distinction. Our framework relies on the availability of key process characteristics. Therefore, Section 3 classifies and provides examples of these characteristics, along with showing how to extract them from raw event data. Section 4 presents event-selection filters to determine the instances of the learning problem. In Section 5 the filters and the process characteristics are used to provide an overview of the wide range of questions that can be answered using our framework. Several well-studied problems turn out to be specific instances of the more general problem considered in this paper. Section 6 discusses the notion of clustering parts of the log based on prediction trees. Section 7 illustrates the implementation of the framework as well as two case studies that have benefitted from the application of the framework. Finally, Section 8 positions

[1] See the *FeaturePrediction* package available in ProM 6.4 or newer, downloadable from http://www.promtools.org.