Chapter 2.19
An Information-Theoretic Framework for Process Structure and Data Mining

Gianluigi Greco
University of Calabria, Italy

Antonella Guzzo
University of Calabria, Italy

Luigi Pontieri
Institute of High Performance Computing and Networks, Italy

INTRODUCTION

Process mining is a key technology for advanced business process management, aimed at supporting the (re)design phase of process-oriented systems like workflow management (WFM), enterprise resource planning (ERP), customer relationship management (CRM), business to business (B2B), and supply chain management (SCM) systems. In fact, based on the log data—a.k.a. transactional log or audit trail—that are gathered by these systems, process mining techniques (Herbst & Karagiannis, 2000; Schimm, 2003; van der Aalst et al., 2003; van der Aalst, Weijters, & Maruster, 2004) are designed to discover the underlying process model and constraints explaining the episodes recorded. The "mined" model provides the users with a syntectic view of the operations involved in the process, which can be the first step leading to supporting the process with a workflow system. Even if the process is already equipped with a workflow schema, such a model can profitably help in re-engineering that schema.

While the output of process mining techniques has been originally defined to be a *unique* model, recent research (Greco, Guzzo, Pontieri, & Saccà, 2006) has argued the importance of explicitly singling out the variants (i.e., the use cases) for the process at hand. This is particularly useful in the case of a complex process, possibly involving hundreds of activities, for which a unique schema may be an overly-detailed and inappropriate description of the actual process behavior, for it mixes semantically different scenarios. Technically, different variants of a process can be detected by partitioning the traces into clusters so that a different schema is eventually induced for each cluster, by way of traditional process mining algorithms.

However, despite the efforts spent in designing process mining techniques, the actual impact of such techniques in the industry is endangered by some simplifying assumptions. In particular, most of the approaches in the literature, and specifically the ones addressed to the mining of different process variants are *propositional* in that in order to extract a model for the process, they only take account for the sequence of task identifiers associated with logged instances, thereby completely disregarding all *non-structural* information kept by many real systems such as activity executors, time-stamps, parameter values, and various performance data.

Contributions

In this article, a further step toward enhancing the process mining framework is made by presenting an approach to discovering variants for a process by means of a technique for clustering log traces, which takes care of both structural and non-structural aspects. Specifically, beside the list of activity identifiers, each trace is also equipped with a number of *metrics*, which are meant to characterize some performance measures for the enactment of the process such as the total processing time and the quality of the process. These measures are very relevant from a business process intelligence perspective; in fact, the need of getting explanations for why a certain metric has a certain value and for predicting the value of such metrics in forthcoming executions has clearly emerged in advanced frameworks geared for the industry (Casati, Castellanos, Dayal, & Shan, 2005).

In order to take care of the different (both structural and not-structural) execution facets in the clustering, we introduce and discuss an information-theoretic framework that extends previous formalizations in Dhillon, Mallela, & Modha (2003) and Berkhin & Becher (2002), in a way that the structural information as well as each of the performance measures is represented by a proper domain, which is correlated to the

Figure 1. A workflow schema for process HANDLEORDER