

Process Mining in the Large: Preprocessing, Discovery, and Diagnostics

Abstract

Process mining techniques attempt to extract non-trivial process related knowledge and interesting insights from event logs. Process mining has made significant progress in less than a decade since its inception. Today, we are at the cross roads of an increasing number of unprecedented domains and new applications willing to apply and adopt process mining. Process mining is being looked at even in applications that are atypical of workflow systems. Analysis of event logs of high-tech systems such as X-Ray machines and CT scanners (medical systems), copiers and printers, mission-critical information systems etc., are a few examples to quote. Existing process mining techniques have shown their applicability in workflow-like processes. However, analyzing less structured processes as seen in the context of high-tech systems is more difficult e.g., traditional process discovery algorithms yield spaghetti-like process models that are hard to comprehend. Event logs from high-tech systems bring along with it a new set of challenges to cope with e.g., fine granular events, voluminous data, flexibility of systems and heterogeneity of cases etc. *This calls for a need for additional techniques and approaches to augment the repertoire of process mining techniques.*

We advocate that the problems arising in analyzing large scale event logs can be tackled from two fronts viz., through *event log simplification* and *advancements in process mining*. We address both these directions in this thesis. We develop an approach to form *abstractions* over events by exploiting the common execution patterns manifested in an event log. Event logs can be simplified by replacing the low-level events with abstract activities thereby mitigating the problem of fine granular events. Heterogeneity in cases can be dealt by first partitioning an event log into subsets of homogenous cases and analyzing these subsets independently. We propose context-aware approaches to *trace clustering* so that homogenous cases are grouped together and show that this assists in enhancing the insights uncovered. Another issue with contemporary approaches to process mining is that they assume the process to be in a steady state. However, in reality, processes can change due to changing circumstances. We introduce the topic of *concept drift* to deal with process changes and propose techniques to detect points of change in an event log. Detection of such change points enables the selection of cases and putting the analysis results in perspective to process variants.

Though log simplification is a critical step, it alone is not sufficient to address the questions with which process mining is sought after as a means to provide answers to. Process mining being a relatively young field, several challenging problems remain to be addressed e.g., traditional process discovery techniques allow only the discovery of flat models. We take a step forward towards addressing some of these in this thesis. We propose a two-phase approach to process discovery that enables the discovery of hierarchical process models (process maps). The majority of research in process mining so far is devoted to process discovery. Process diagnostics is gaining significance in recent years. Process diagnostics encompasses process performance analysis, anomaly detection, diagnosis (root-cause analysis), inspection of interesting patterns and the like. We propose a means of replaying an event log onto process maps so that process performance can be measured and bottlenecks identified. Further, we develop a trace visualization technique, called as *trace alignment*,

that aligns the traces in such a way that event logs can be explored easily. We show that trace alignment is extremely helpful when analyzing event logs from less-structured processes and has significant promise in process diagnostics. We also propose techniques for discovering patterns of behavior that discriminate between categories of cases e.g., normal and fraudulent insurance claims.

The work presented in this thesis has been evaluated on real-life case studies and is supported by concrete implementations integrated in the ProM framework.