Process Mining Put Into Context

Wil M.P. van der Aalst[1,2] and Schahram Dustdar[3]

[1] Eindhoven University of Technology
[2] Queensland University of Technology
[3] Technical University of Vienna

Abstract. Process mining techniques can be used to discover and analyze business processes based on raw event data. This article first summarizes guiding principles and challenges taken from the recently released Process Mining Manifesto. Next, the authors argue that the *context* in which events occur should be taken into account when analyzing processes. Contextualized event data can be used to extend the scope of process mining and improve the quality of analysis results.

Process mining is an emerging research discipline that sits between computational intelligence and data mining on the one hand, and process modeling and analysis on the other hand [2]. Starting point for process mining is an *event log*. All process mining techniques assume that it is possible to *sequentially* record *events* such that each event refers to an *activity* (i.e., a well-defined step in the process) and is related to a particular *case* (i.e., a process instance). Event logs may store additional information such as the *resource* (i.e., person or device) executing or initiating an activity, the *timestamp* of an event, or *data elements* recorded with an event (e.g., the size of an order). Event logs can be used to discover, monitor and improve processes based on facts rather than fiction. There are three types of process mining.

- *Discovery*: take an event log and produce a model without using any other a-priori information. There are dozens of techniques to extract a process model from raw event data. For example, the classical α algorithm is able to discover a Petri net by identifying basic process patterns in an event log [3]. For many organizations it is surprising to see that existing techniques are indeed able to discover real processes based on merely example executions recorded in event logs. Process discovery is often used as a starting point for other types of analysis.
- *Conformance*: an existing process model is compared with an event log of the same process. The comparison shows where the real process deviates from the modeled process. Moreover, it is possible to quantify the level of conformance and differences can be diagnosed. Conformance checking can be used to check if reality, as recorded in the log, conforms to the model and vice versa. There are various applications for this (compliance checking, auditing, six-sigma, etc.) [2].

– *Enhancement*: take an event log and process model and extend or improve the model using the observed events. Whereas conformance checking measures the alignment between model and reality, this third type of process mining aims at changing or extending the a-priori model. For instance, by using timestamps in the event log one can extend the model to show bottlenecks, service levels, throughput times, and frequencies [2].

Over the last decade, event data have become readily available and process mining techniques have matured. Moreover, managements trends related to process improvement (e.g., Six Sigma, TQM, CPI, and CPM) and compliance (SOX, BAM, etc.) can benefit from process mining. Process mining has become one of the "hot topics" in Business Process Management (BPM) research and there is considerable interest from industry in process mining. More and more software vendors started adding process mining functionality to their tools.

IEEE Task Force on Process Mining

The growing interest in log-based process analysis motivated the establishment of the *IEEE Task Force on Process Mining*. The goal of this task force is to promote the research, development, education and understanding of process mining. The task force was established in 2009 in the context of the Data Mining Technical Committee of the Computational Intelligence Society of the IEEE. Members of the task force include representatives of more than a dozen commercial software vendors (e.g., Pallas Athena, Software AG, Futura Process Intelligence, HP, IBM, Fujitsu, Infosys, and Fluxicon), ten consultancy firms (e.g., Gartner and Deloitte) and over twenty universities.
Concrete objectives of the task force are:

– to make end-users, developers, consultants, business managers, and researchers aware of the state-of-the-art in process mining,
– to promote the use of process mining techniques and tools and stimulate new applications,
– to play a role in standardization efforts for logging event data,
– to organize tutorials, special sessions, workshops, panels, and
– to publish articles, books, videos, and special issues of journals.

See http://www.win.tue.nl/ieeetfpm/ for more information about the activities of the task force.

Process Mining Manifesto

The IEEE Task Force on Process Mining recently released a manifesto describing *guiding principles* and *challenges* [5]. The manifesto aims to increase the visibility of process mining as a new tool to improve the (re)design, control, and

support of operational business processes. It is intended to guide software developers, scientists, consultants, and end-users. As an introduction to the state-of-the-art in process mining, we briefly summarize the main findings reported in the manifesto [5].

Table 1. Six guiding principles [5]

GP1	**Event Data Should Be Treated as First-Class Citizens** Events should be *trustworthy*, i.e., it should be safe to assume that the recorded events actually happened and that the attributes of events are correct. Event logs should be *complete*, i.e., given a particular scope, no events may be missing. Any recorded event should have well-defined *semantics*. Moreover, the event data should be *safe* in the sense that privacy and security concerns are addressed when recording the event log.
GP2	**Log Extraction Should Be Driven by Questions** Without concrete questions it is very difficult to extract meaningful event data. Consider, for example, the thousands of tables in the database of an ERP system like SAP. Without questions one does not know where to start.
GP3	**Concurrency, Choice and Other Basic Control-Flow Constructs Should be Supported** Basic workflow *patterns* supported by all mainstream languages (e.g., BPMN, EPCs, Petri nets, BPEL, and UML activity diagrams) are *sequence*, *parallel routing* (AND-splits/joins), *choice* (XOR-splits/joins), and *loops*. Obviously, these patterns should be supported by process mining techniques.
GP4	**Events Should Be Related to Model Elements** Conformance checking and enhancement heavily rely on the relationship between *elements in the model* and *events in the log*. This relationship may be used to "replay" the event log on the model. Replay can be used to reveal discrepancies between event log and model (e.g., some events in the log are not possible according to the model) and can be used to enrich the model with additional information extracted from the event log (e.g., bottlenecks are identified by using the timestamps in the event log).
GP5	**Models Should Be Treated as Purposeful Abstractions of Reality** A model derived from event data provides a *view on reality*. Such a view should serve as a purposeful abstraction of the behavior captured in the event log. Given an event log, there may be multiple views that are useful.
GP6	**Process Mining Should Be a Continuous Process** Given the dynamical nature of processes, it is not advisable to see process mining as a one-time activity. The goal should not be to create a fixed model, but to breathe life into process models such that users and analysts are encouraged to look at them on a daily basis.

Guiding Principles

As with any new technology, there are obvious mistakes that can be made when applying process mining in real-life settings. Therefore, the six guiding princi-

ples listed in Table 1 aim to prevent users/analysts from making such mistakes. As an example, consider guiding principle *GP4*: "Events Should Be Related to Model Elements". It is a misconception that process mining is limited to control-flow discovery, other perspectives such as the organizational perspective, the time perspective, and the data perspective are equally important. However, the control-flow perspective (i.e., the ordering of activities) serves as the layer connecting the different perspectives. Therefore, it is important to relate events to activities in the model. Conformance checking and model enhancement heavily rely on this relationship. After relating events to model elements, it is possible to "replay" the event log on the model [2]. Replay may be used to reveal discrepancies between an event log and a model, e.g., some events in the log are not possible according to the model. Techniques for conformance checking quantify and diagnose such discrepancies. Timestamps in the event log can be used to analyze the temporal behavior during replay. Time differences between causally related activities can be used to add average/expected waiting times to the model. These examples illustrate the importance of guiding principle GP4; the relation between events in the log and elements in the model serves as a starting point for different types of analysis.

Challenges

Process mining is an important tool for modern organizations that need to manage non-trivial operational processes. On the one hand, there is an incredible growth of event data. On the other hand, processes and information need to be aligned perfectly in order to meet requirements related to compliance, efficiency, and customer service. Despite the applicability of process mining there are still important challenges that need to be addressed; these illustrate that process mining is an emerging discipline. Table 2 lists the eleven challenges described in the Process Mining Manifesto [5]. As an example consider Challenge *C4*: "Dealing with Concept Drift". The term *concept drift* refers to the situation in which the process is changing while being analyzed [4]. For instance, in the beginning of the event log two activities may be concurrent whereas later in the log these activities become sequential. Processes may change due to periodic/seasonal changes (e.g., "in December there is more demand" or "on Friday afternoon there are fewer employees available") or due to changing conditions (e.g., "the market is getting more competitive"). Such changes impact processes and it is vital to detect and analyze them. However, most process mining techniques analyze processes as if they are in steady-state [4].

Using a Broader Context

Processes are executed in a particular *context*, but this context is often neglected during analysis [6,7]. We distinguish four types of contexts: (a) *instance* context, (b) *process* context, (c) *social* context, and (d) *external* context. Existing process mining techniques tend to use a rather narrow context, i.e., only the instance

Table 2. Some of the most important process mining challenges identified in the manifesto [5]

C1	**Finding, Merging, and Cleaning Event Data** When extracting event data suitable for process mining several challenges need to be addressed: data may be *distributed* over a variety of sources, event data may be *incomplete*, an event log may contain *outliers*, logs may contain events at *different level of granularity*, etc.
C2	**Dealing with Complex Event Logs Having Diverse Characteristics** Event logs may have very different characteristics. Some event logs may be extremely large making them difficult to handle whereas other event logs are so small that not enough data is available to make reliable conclusions.
C3	**Creating Representative Benchmarks** Good benchmarks consisting of example data sets and representative quality criteria are needed to compare and improve the various tools and algorithms.
C4	**Dealing with Concept Drift** The process may be changing while being analyzed. Understanding such concept drifts is of prime importance for the management of processes.
C5	**Improving the Representational Bias Used for Process Discovery** A more careful and refined selection of the representational bias is needed to ensure high-quality process mining results.
C6	**Balancing Between Quality Criteria such as Fitness, Simplicity, Precision, and Generalization** There are four competing quality dimensions: (a) fitness, (b) simplicity, (c) precision, and (d) generalization. The challenge is to find models that score good in all four dimensions.
C7	**Cross-Organizational Mining** There are various use cases where event logs of multiple organizations are available for analysis. Some organizations work together to handle process instances (e.g., supply chain partners) or organizations are executing essentially the same process while sharing experiences, knowledge, or a common infrastructure. However, traditional process mining techniques typically consider one event log in one organization.
C8	**Providing Operational Support** Process mining is not restricted to off-line analysis and can also be used for online operational support. Three operational support activities can be identified: *detect*, *predict*, and *recommend*.
C9	**Combining Process Mining With Other Types of Analysis** The challenge is to combine automated process mining techniques with other analysis approaches (optimization techniques, data mining, simulation, visual analytics, etc.) to extract more insights from event data.
C10	**Improving Usability for Non-Experts** The challenge is to hide the sophisticated process mining algorithms behind user-friendly interfaces that automatically set parameters and suggest suitable types of analysis.
C11	**Improving Understandability for Non-Experts** The user may have problems understanding the output or is tempted to infer incorrect conclusions. To avoid such problems, the results should be presented using a suitable representation and the trustworthiness of the results should always be clearly indicated.

in isolation is considered. However, the handling of cases is influenced by a much broader context. Therefore, analysis should not abstract from anything not directly related to the individual instance.

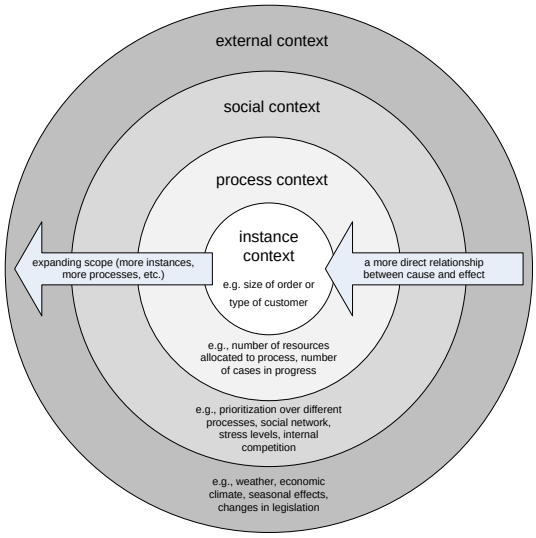

Fig. 1. Different levels of context data that may influence the process that is analyzed using process mining.

Instance Context

Process instances (i.e., cases) may have various properties that influence their execution. Consider for example the way a customer order is handled. The type of customer placing the order may influence the path the instance follows in the process. The size of the order may influence the type of shipping selected or may influence the transportation time. These properties can be directly related to the individual process instance and we refer to them as the *instance context*. Typically, it is not difficult to discover relationships between the instance context and the observed behavior of the case. For example, one could discover that an activity is typically skipped for gold customers.

Process Context

A process may be instantiated many times, e.g., thousands of customer orders are handled by the same process per year. Yet, the corresponding process model typically describes the life-cycle of one order in isolation. Although interactions among instances are not made explicit in such models, they may influence each

other. For example, instances may compete for the same resources. An order may be delayed by too much work-in-progress. Looking at one instance in isolation is not sufficient for understanding the observed behavior. Process mining techniques should also consider the *process context*, e.g., the number of instances being handled and the number of resources available for the process. For example, when predicting the expected remaining flow time for a particular case one should not only consider the instance context (e.g., the status of the order) but also the process context (e.g., workload and resource availability).

Social Context

The process context considers all factors that can be directly related to a process and its instances. However, people and organizations are typically not allocated to a single process and may be involved in many different processes. Moreover, activities are executed by people that operate in a social network. Friction between individuals may delay process instances and the speed at which people work may vary due to circumstances that cannot be fully attributed to the process being analyzed. All of these factors are referred to as the *social context*. This context characterizes the way in which people work together *within a particular organization*. Today's process mining techniques tend to neglect the social context even though it is clear that this context directly impacts the way that cases are handled.

How people work

When using existing mainstream business process modeling languages, it is only possible to describe human resources in a very naive manner. Often people are involved in many different processes, e.g., a manager, doctor, or specialist may perform tasks in a wide range of processes. Seen from the viewpoint of a single process, these individuals may have a very low utilization. However, a manager that needs to distribute her attention over dozens of processes may easily become a bottleneck. However, when faced with unacceptable delays the same manager can also decide to devote more attention to the congested process and quickly resolve all problems. Related is the so-called "Yerkes-Dodson Law of Arousal" that describes the phenomenon that people work at different speeds based on their workload. Not just the distribution of attention over various processes matters: also the workload-dependent working speeds determine the effective resource capacity for a particular process [1].

External Context

The *external context* captures all factors that are part of an even wider ecosystem that extends beyond the control sphere of the organization. For example,

the weather, the economic climate, and changing regulations may influence the way that cases are being handled. The weather may influence the workload, e.g., a storm or flooding may lead to increased volumes of insurance claims. Changing oil prices may influence the number of customer orders (e.g., the demand for heating oil increases when prices drop). More stringent identity checks may influence the order in which social security related activities are being executed. Although the external context can have a dramatic impact on the process being analyzed, it is difficult to select the relevant variables. Learning the effect of the external context is closely related to the identification of concept drift, e.g., a process may gradually change due to external seasonal effects.

Curse of Dimensionality

The four types of context described in this article describe a continuum of factors that may influence a process. The factors closely related to a process instance are easy to identify. However the social and external contexts are difficult to capture in a few variables that can be used by process mining algorithms. Moreover, we are often faced with the so-called "Curse of Dimensionality", i.e., in high-dimensional feature spaces enormous amounts of event data are required to reliably learn the effect of contextual factors. Therefore, additional research is needed to "put process mining in context".

References

1. W.M.P. van der Aalst. Business Process Simulation Revisited. In J. Barjis, editor, *Enterprise and Organizational Modeling and Simulation*, volume 63 of *Lecture Notes in Business Information Processing*, pages 1–14. Springer-Verlag, Berlin, 2010.
2. W.M.P. van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer-Verlag, Berlin, 2011.
3. W.M.P. van der Aalst, A.J.M.M. Weijters, and L. Maruster. Workflow Mining: Discovering Process Models from Event Logs. *IEEE Transactions on Knowledge and Data Engineering*, 16(9):1128–1142, 2004.
4. R.P. Jagadeesh Chandra Bose, W.M.P. van der Aalst, I. Zliobaite, and M. Pechenizkiy. Handling Concept Drift in Process Mining. In H. Mouratidis and C. Rolland, editors, *International Conference on Advanced Information Systems Engineering (Caise 2011)*, volume 6741 of *Lecture Notes in Computer Science*, pages 391–405. Springer-Verlag, Berlin, 2011.
5. IEEE Task Force on Process Mining. Process Mining Manifesto. In *BPM Workshops*, Lecture Notes in Business Information Processing. Springer-Verlag, Berlin, 2011.
6. K. Ploesser, M. Peleg, P. Soffer, M. Rosemann, and J. Recker. Learning from Context to Improve Business Processes. *BPTrends*, pages 1–7, January 2009.
7. M. Rosemann, J. Recker, and C. Flender. Contextualisation of Business Processes. *International Journal of Business Process Integration and Management*, 3(1):47–60, 2008.