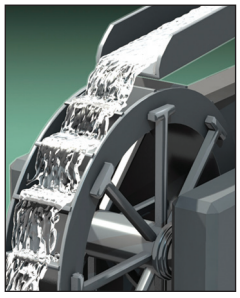
Editor: Schahram Dustdar • dustdar@infosys.tuwien.ac.at



Process Mining Put into Context

Wil M.P. van der Aalst • *Eindhoven University of Technology*

Schahram Dustdar • *Technical University of Vienna*

Process mining techniques help organizations discover and analyze business processes based on raw event data. The recently released "Process Mining Manifesto" presents guiding principles and challenges for process mining. Here, the authors summarize the manifesto's main points and argue that analysts should take into account the context in which events occur when analyzing processes.

Process mining is an emerging research discipline that sits between computational intelligence and data mining on one hand, and process modeling and analysis on the other.[1] Process mining techniques help organizations discover and analyze business processes based on raw event data.

As event data have become readily available over the past decade, process mining techniques have matured. Moreover, management trends related to process improvement (such as Six Sigma, TQM, CPI, and CPM) and compliance (SOX, BAM, and so on) can benefit from such techniques. Process mining has become a "hot topic" in business process management (BPM) research, and industry has expressed considerable interest in employing it as well. More and more software vendors are adding process mining functionality to their tools.

Here, we briefly explain process mining and delve into some of the principles and challenges described in the recently released "Process Mining Manifesto."[2] Moreover, we argue that the *context* should be taken into account when analyzing processes.

Process Mining Basics

The starting point for process mining is an event log. All process mining techniques assume that events can be sequentially recorded such that each event refers to an activity (that is, a well-defined step in the process) and is related to a particular case (a process instance). Event logs might store additional information such as the resource (person or device) executing or initiating an activity, an event's time stamp, or data elements recorded with an event (such as the size of an order). Organizations can use event logs to discover, monitor, and improve processes based on facts rather than fiction. Three types of process mining exist:

- *Discovery* takes an event log and produces a model without using any other a priori information. Dozens of techniques exist for extracting a process model from raw event data. The classical α algorithm, for example, can discover a Petri net by identifying basic process patterns in an event log.[3] Many organizations are surprised to see that existing techniques can indeed discover real processes based merely on example executions recorded in event logs. Organizations thus often use process discovery as a starting point for other types of analysis.
- *Conformance* compares an existing process model with the same process's event log. This comparison shows where the real process deviates from the modeled one. Moreover, process mining techniques can quantify the level of conformance and diagnose differences. Conformance checking can determine if reality, as recorded in the log, conforms to the model

 Published by the IEEE Computer Society 1089-7801/12/$31.00 © 2012 IEEE

The 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. Its goal is to promote the research, development, education, and understanding of this technology. The task force was established in 2009 in the context of the IEEE Computational Intelligence Society's Data Mining Technical Committee. Task force members include representatives from more than a dozen commercial software vendors (including Pallas Athena, Software AG, Futura Process Intelligence, Hewlett-Packard, IBM, Fujitsu, Infosys, and Fluxicon), 10 consultancy firms (such as Gartner and Deloitte), and more than 20 universities.

The task force's concrete objectives are to

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

More information about the task force's activities is available at www.win.tue.nl/ieeetfpm.

and vice versa. Such functionality supports compliance checking, auditing, Six Sigma, and so on.[1]

- *Enhancement* takes an event log and process model and extends or improves the model using the observed events. Whereas conformance checking measures the alignment between model and reality, this type of process mining changes or extends the a priori model. For instance, by using time stamps in the event log, process mining tools can extend a model to show bottlenecks, service levels, throughput times, and frequencies.[1]

Process Mining Manifesto

The IEEE Task Force on Process Mining (see the sidebar) recently released a manifesto describing guiding principles and challenges.[2] The manifesto aims to increase process mining's visibility as a new tool for improving the (re)design, control, and support of operational business processes. It's intended to guide software developers, scientists, consultants, and end users in adopting this technology. Let's look briefly at this manifesto's main findings.

Guiding Principles

As with any new technology, people can make mistakes when applying process mining in real-life settings. The six guiding principles Table 1 lists aim to prevent users and analysts from making such mistakes. Consider GP4: events should be related to model elements. It's a misconception that process mining is limited to control-flow discovery; other perspectives, such as organizational, time, and data perspectives, are equally important. However, the control flow (that is, the order of activities) serves as the layer connecting the different perspectives. So, events must be related to activities in the model. Conformance checking and model enhancement rely heavily on this relationship. After relating events to model elements, process mining tools can "replay" the event log on the model.[1] Replay can reveal discrepancies between an event log and a model, and techniques for conformance checking quantify and diagnose such discrepancies. Time stamps in the event log can help analyze temporal behavior during replay. Time differences between causally related activities can add average and expected wait times to the model. These examples illustrate GP4's importance; the relationship 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 must manage nontrivial operational processes. On one hand, the volume of event data is growing exponentially. On the other hand, processes and information must align perfectly if organizations are to meet compliance, efficiency, and customer service requirements. Despite process mining's applicability in such circumstances, important challenges remain and illustrate that it's an emerging discipline. Table 2 lists the 11 challenges described in the "Process Mining Manifesto."[2] Consider C4: dealing with concept drift. The term *concept drift* refers to a situation in which the process is changing while being analyzed.[4] For example, at the beginning of an event log, two activities might be concurrent, whereas later in the log they become sequential. Processes can change due to periodic or seasonal changes ("in December, there is more demand" or "on Friday afternoon, fewer employees are available") or to changing conditions ("the market is getting more competitive"). Such changes affect processes, and organizations must detect and analyze them. However, most process mining techniques analyze processes as if they're in a steady state.[4]

Using a Broader Context

Organizations execute processes in a particular context that's often neglected during analysis.[5,6] We distinguish

Table 1. The six guiding principles of the "Process Mining Manifesto."[2]

Guiding principle	Characteristics
G1. Event data should be treated as first-class citizens.	Events should be trustworthy — that is, we should be able to safely assume that recorded events actually happened and that their attributes are correct. Event logs should be complete; given a particular scope, no events should be missing. Any recorded event should have well-defined semantics. Moreover, event data should be safe in the sense that privacy and security concerns are addressed when the event log is recorded.
G2. Log extraction should be driven by questions.	Without concrete questions, extracting meaningful event data is very difficult. Consider, for example, the thousands of tables in the database of an enterprise resource planning (ERP) system such as SAP. Without questions, we don't know where to start.
G3. Concurrency, choice, and other basic control-flow constructs should be supported.	Basic workflow patterns supported by all mainstream languages (such as the Business Process Modeling Notation, event-driven process chains, Petri nets, the Business Process Execution Language, and UML activity diagrams) include sequence, parallel routing (AND splits/joins), choice (XOR splits/joins), and loops. Obviously, process mining techniques should support these patterns.
G4. Events should be related to model elements.	Conformance checking and enhancement rely heavily on the relationship between elements in the model and events in the log. This relationship can help "replay" the event log on the model. We can use replay to reveal discrepancies between the event log and model (such as some events in the log being impossible according to the model) and enrich the model with additional information extracted from the log (for example, indentifying bottlenecks using the event log's time stamps).
G5. Models should be treated as purposeful abstractions of reality.	A model derived from event data provides a view on reality that should serve as a purposeful abstraction of the behavior the event log captures. Given a single event log, multiple views might be useful.
G6. Process mining should be a continuous process.	Given processes' dynamic nature, viewing process mining as a one-time activity is inadvisable. The goal shouldn't be to create a fixed model but rather to breathe life into process models such that users and analysts are encouraged to look at them daily.

four types of context: *instance*, *process*, *social*, and *external* (see Figure 1). Existing process mining techniques tend to use a rather narrow context — that is, only the process instance itself is considered. However, a much broader context influences the way in which instances are handled; analysis shouldn't abstract from anything not directly related to the individual instance.

Instance Context

Process instances (that is, cases) might have various properties that influence their execution. Consider the way businesses handle a customer order. The type of customer placing the order can influence the path the instance follows in the process. The order's size can influence the type of shipping the customer selects or the transportation time. These properties can directly relate to the individual process instance; we refer to them as the instance context. Typically, discovering relationships between the instance context and the case's observed behavior isn't difficult. We might, for example, discover that an activity is typically skipped for VIP customers.

Process Context

A process might be instantiated many times — for example, the process can handle thousands of customer orders per year. Yet, the corresponding process model typically describes one order's life cycle in isolation. Although interactions among instances aren't made explicit in such models, they can influence each other. Instances might compete for the same resources, and an order might be delayed by too much work-in-progress. Looking at one instance in isolation isn't sufficient for understanding the observed behavior. Process mining techniques should also consider the process context, such as the number of instances being handled and resources available for the process. When predicting the expected remaining flow time for a particular case, for example, the analysis tool should consider not only the order's status (instance context) but also the workload and resource availability (process context).

Social Context

The process context considers all factors directly related to a process and its instances. However, people and organizations typically aren't allocated to a single process and might be involved in many different processes. Moreover, activities are executed by people operating in a social network. Friction between individuals can delay process instances, and the speed at which people work might vary due to circumstances that aren't fully attributable

Table 2. Process mining challenges identified in the manifesto.[2]

Challenges	Characteristics
C1. Finding, merging, and cleaning event data	Extracting event data suitable for process mining presents several challenges: data might be distributed over various sources, event data might be incomplete, an event log might contain outliers, or logs might contain events at different levels of granularity.
C2. Dealing with complex event logs with diverse characteristics	Event logs can have very different characteristics. Some might be extremely large, making them difficult to handle, whereas others are so small that not enough data is available to make reliable conclusions.
C3. Creating representative benchmarks	Good benchmarks consisting of example datasets and representative quality criteria are needed to compare and improve various process mining tools and algorithms.
C4. Dealing with concept drift	The process might change as it's being analyzed. Understanding such concept drifts is highly important for process management.
C5. Improving the representational bias used for process discovery	A more careful and refined selection of representational bias is required to ensure high-quality process mining results.
C6. Balancing between quality criteria	Process mining has four competing quality dimensions: fitness, simplicity, precision, and generalization. The challenge is finding models that score well in all four dimensions.
C7. Cross-organizational mining	In various use cases, multiple organizations' event logs are available for analysis. Some organizations work together to handle process instances (for instance, supply chain partners), whereas others execute 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 isn't restricted to offline analysis; it can also provide online operational support. Three operational support activities are 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, and so on) to extract more insights from event data.
C10. Improving usability for nonexperts	This challenge is to hide sophisticated process mining algorithms behind user-friendly interfaces that automatically set parameters and suggest suitable types of analysis.
C11. Improving understandability for nonexperts	Users might have problems understanding the output or be tempted to infer incorrect conclusions. To avoid such problems, process mining results should use a suitable representation and always clearly indicate their trustworthiness.

to the process being analyzed (see the "How People Work" sidebar). We refer to all these factors as the social context, which characterizes how people work together within a particular organization. Today's process mining techniques tend to neglect the social context, even though it directly impacts how people and organizations handle cases.

External Context

The external context captures factors that are part of an ecosystem that extends beyond an organization's control sphere. For example, the weather, the economic climate, and changing regulations might influence how organizations handle cases. The weather might influence the workload, as when a storm or flooding leads to increased insurance claims. Changing oil prices can influence customer orders, as when the demand for heating oil increases as prices drop. More stringent identity checks influence the order in which a government organization executes social-security-related activities. Although external context can have a dramatic impact on the process being analyzed, selecting relevant variables is difficult. Learning the external context's effects is closely related to identifying concept drift — for example, a process might gradually change due to external seasonal effects.

The four context types we describe demonstrate a continuum of factors that can influence a process. The factors closely related to a process instance are easy to identify.However, social and external contexts are difficult to capture in a few variables that process mining algorithms can use. Moreover, we're often faced with the so-called "curse of dimensionality" — that is, in high-dimensional feature spaces, enormous amounts of event data are required to reliably learn contextual factors' effects. Additional research is needed before we can "put process mining in context."

How People Work

When using existing mainstream business process modeling languages, we can only describe human resources in a very naïve manner. People are often involved in many different processes; a manager, doctor, or specialist might perform tasks in a wide range of processes. Seen from a single process viewpoint, these individuals might have a very low utilization. However, a manager who needs to distribute his or her attention over dozens of processes might easily become a bottleneck. 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," which describes the phenomenon that people work at different speeds based on their workload. Not only the distribution of attention over various processes matters: workload-dependent working speeds also determine the effective resource capacity for a particular process.[1]

Reference

1. W.M.P. van der Aalst, "Business Process Simulation Revisited," *Enterprise and Organizational Modeling and Simulation*, Lecture Notes in Business Information Processing, vol. 63, J. Barjis, ed., Springer, 2010, pp. 1–14.

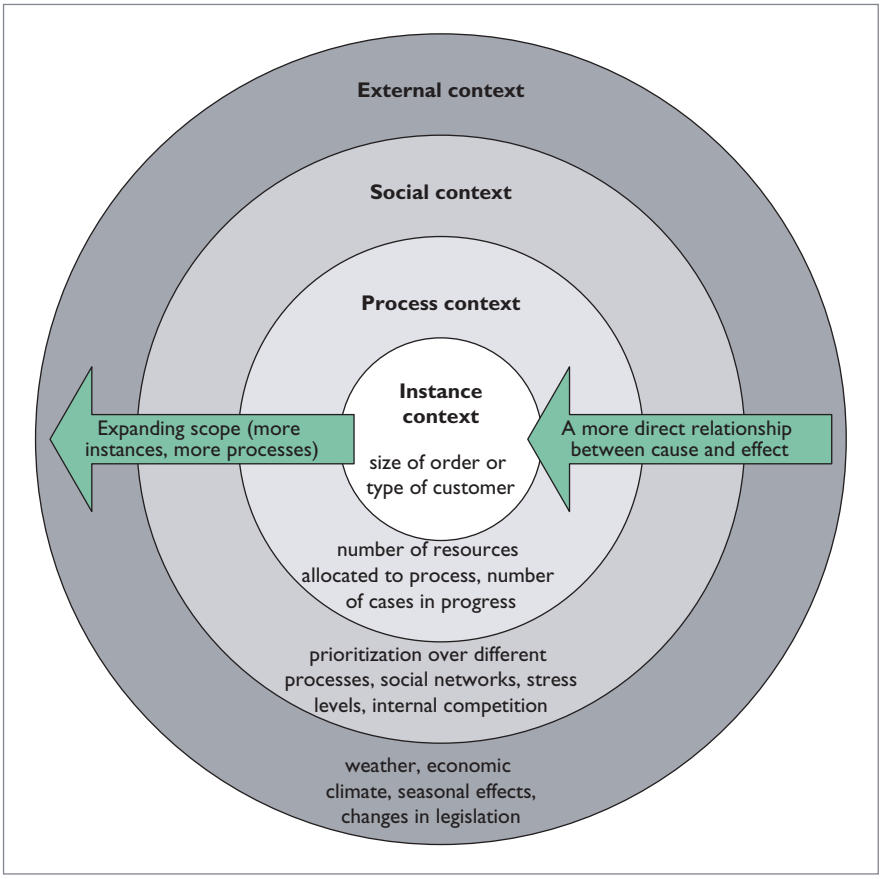


Figure 1. Levels of context data. Context can influence processes, but is often neglected during process mining analysis.

References

1. W.M.P. van der Aalst, *Process Mining: Discovery, Conformance, and Enhancement of Business Processes*, Springer, 2011.
2. IEEE Task Force on Process Mining, "Process Mining Manifesto," *Proc. Business Process Mining Workshops*, Lecture Notes in Business Information Processing, Springer, 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 Trans. Knowledge and Data Eng.*, vol. 16, no. 9, 2004, pp. 1128–1142.
4. R.P. Jagadeesh Chandra Bose et al., "Handling Concept Drift in Process Mining," *Proc. Int'l Conf. Advanced Information Systems Eng.* (CAiSE 11), LNCS 6741, Springer, 2011, pp. 391–405.
5. K. Ploesser et al., "Learning from Context to Improve Business Processes," *BPTrends*, Jan. 2009, pp. 1–7.
6. M. Rosemann, J. Recker, and C. Flender, "Contextualization of Business Processes," *Int'l J. Business Process Integration and Management*, vol. 3, no. 1, 2008, pp. 47–60.

Wil M.P. van der Aalst is a full professor at Eindhoven University of Technology and part-time professor at Queensland University of Technology. His research interests include process modeling, business process management, process mining, and concurrency. Van der Aalst has a PhD in mathematics from Eindhoven University of Technology. He wrote the first book on process mining and chairs the IEEE Task Force on Process Mining. Contact him at w.m.p.v.d.aalst@tue.nl.

Schahram Dustdar is a full professor of computer science (informatics) with a focus on Internet technologies and heads the Distributed Systems Group, Institute of Information Systems, at the Vienna University of Technology (TU Wien). Dustdar is an ACM Distinguished Scientist. Contact him at dustdar@infosys.tuwien.ac.at; www.infosys.tuwien.ac.at/.