Challenges in Business Process Mining

W.M.P. van der Aalst

[1] Department of Mathematics and Computer Science,
Eindhoven University of Technology
P.O. Box 513, NL-5600 MB, Eindhoven, The Netherlands.
w.m.p.v.d.aalst@tue.nl
[2] Business Process Management Group, Queensland University of Technology
P.O. Box 2434, Brisbane Qld 4001, Australia.

Abstract. Information systems are becoming more and more intertwined with the operational processes in most organizations. As a result, a multitude of events are recorded by today's information systems. The goal of *process mining* is to use these event data to extract process-related information, e.g., to automatically discover a process model by observing events recorded by some enterprise system. This invited contribution highlights three main research challenges in process mining. Moreover, it discusses the need for more rigorous approaches to evaluating the performance of process mining algorithms. By addressing these challenges and by adapting better evaluation procedures, the maturity of process mining as a research discipline will increase.

Keywords: Process mining, process discovery, business process management, business intelligence.

1 Introduction

The goal of *process mining* is to extract process-related information from event logs, e.g., to automatically discover a process model by observing events recorded by some information system [1, 2]. In the context of ProM we have been working on various process mining techniques [3]. The ProM framework has been developed as a completely plug-able open source environment and groups in the Netherlands, Germany, Italy, France, Austria, Australia, China, Spain, Portugal, and Brazil have contributed to ProM. Currently, ProM is the only comprehensive framework covering the whole process mining spectrum while supporting a wide range of process mining techniques. Thus far we have applied process mining in more than 100 organizations: municipalities (e.g., Alkmaar, Heusden, Harderwijk, etc.), government agencies (e.g., Rijkswaterstaat, Centraal Justitieel Incasso Bureau, Justice department), insurance related agencies (e.g., UWV, ABP), banks (e.g., ING Bank), hospitals (e.g., AMC hospital, Catharina hospital), multinationals (e.g., DSM, Deloitte), high-tech system manufacturers and their customers (e.g., Philips Healthcare, ASML, Thales), and media companies (e.g. Winkwaves). Moreover, several of the ideas developed in the context of the

open source software platform ProM have been adopted in various commercial Business Process Management (BPM) products (e.g., BPM|one, Futura Reflect, and ARIS PPM). This paper is based on these experiences and the state-of-the-art in process mining. In the first part of the paper we highlight three important challenges. First of all, we want to discover *"business process maps"* that are as good as the geographic maps made by cartographers. Second, we want to support tomorrow's auditor by providing *"business process forensics"* based on process mining. Finally, we want to provide organizations with *"TomTom-like navigation functionality"* based on process mining. To address the above three challenges new scientific and technological breakthroughs are needed. Moreover, a review of current literature shows that process mining – despite its potential – is not yet a mature research discipline. Therefore, we conclude this paper with some suggestions for improving the maturity level.

2 Challenges

Process mining starts from the events stored in information systems (e.g., transaction logs, audit trails, databases, message logs) and descriptive or normative process models. More and more information about (business) processes is recorded by information systems in the form of so-called "event logs". IT systems are becoming more and more intertwined with the processes they support, resulting in an "explosion" of available data that can be used for analysis purposes. Unfortunately, *events are typically not recorded in a unified manner and the available data are not exploited well.*

Within large organizations there are also large collections of process models. Processes are modeled for a variety of reasons ranging from process improvement, simulation and certification to corporate governance and workflow automation. However, these *models are often of low quality and not used on a daily basis.* To address these problems *innovative process mining techniques* need to be developed. In particular, we suggest tackling the following three challenges:

- Challenge 1: *Making "business process maps" that are as good as the geographic maps made by cartographers.*
- Challenge 2: *Supporting tomorrow's auditor by providing "business process forensics" based on process mining.*
- Challenge 3: *Providing organizations with "TomTom-like navigation functionality" based on process mining.*

These challenges have in common that they all actively use event data and that process models plays an important role. Moreover, they all refer to weaknesses of today's IT support when managing complex processes. Today's *organizations do not have "maps" of their processes that are actually useful for the people involved in these processes.* Moreover, *basic navigation functionality* that can be found in a €150 TomTom system *is missing.* Contemporary information systems do not provide directions and do not estimate the "arrival time" of work-item and cases. Finally, auditors are not using advanced IT support to analyze the massive

amounts of event data stored in the information system of the organization under review. Instead *they work behind the system's back and need to rely on ad-hoc sample data, human judgment, and simple checks at an aggregate level*. Yet, it is clear that the time is right for introducing "business process forensics" based on event logs.

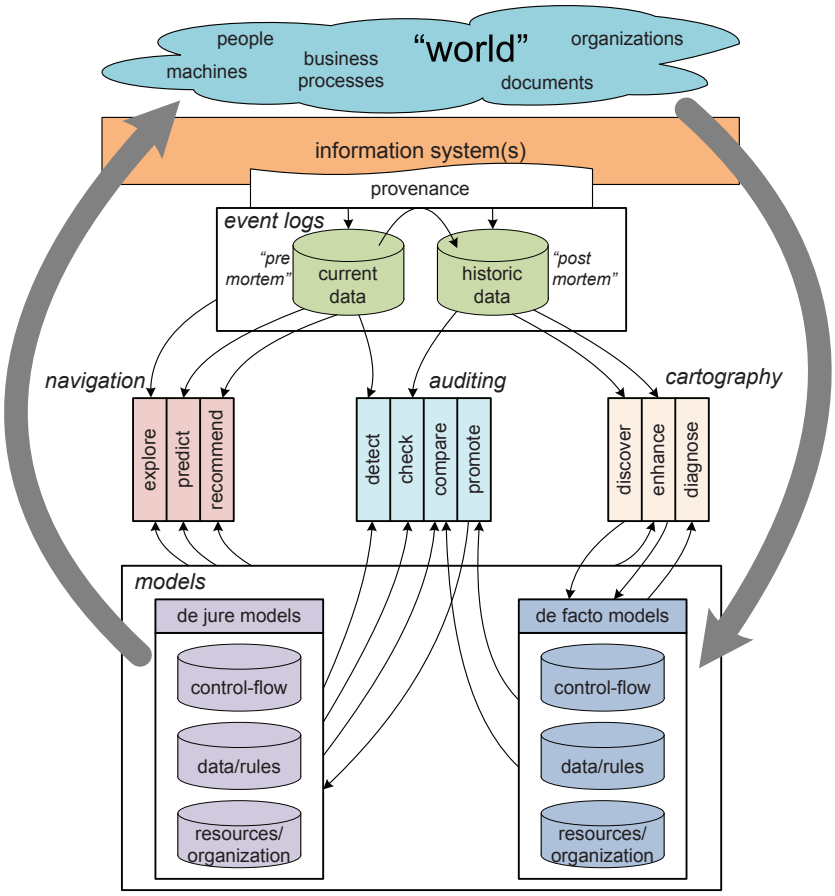

Fig. 1. An overview of the challenges related to business process cartography, auditing, and navigation. Starting point are event logs with "pre mortem" and "post mortem" data. These are related to "de jure" and "de facto" models.

To tackle the three challenges, three main categories of activities have been identified: *cartography*, *auditing*, and *navigation*. These are shown in Figure 1. As mentioned earlier, we assume the existence of a collection of information systems that are supporting a "world" composed of business processes, people, organizations, etc. The *event data* extracted from such systems are the starting

point for process mining. Note that Figure 1 distinguishes between *current data* and *historic data*. The former refers to events of cases (i.e., process instances) that are still actively worked on ("pre mortem"). The latter refers to events of completed cases, i.e., process instances that cannot be influenced anymore ("post mortem"). The lower part of Figure 1 shows two types of models: *de jure models* are normative models that describe a desired or required way of working while *de facto models* aim to describe the actual reality with all of its intricacies (policy violations, inefficiencies, fraud, etc.). Both types of models may cover one or more perspectives and thus describe control-flow, time, data, organization, resource, and/or cost aspects.

In the remainder, we highlight some of the challenges shown in Figure 1.

2.1 Cartography

The first geographical maps data back to the 7th Millennium BC. Since then cartographers have improved their skills and techniques to create maps thereby addressing problems such as clearly representing desired traits, eliminating irrelevant details, reducing complexity, and improving understandability. Today, geographic maps are digital and of high quality. This has fueled innovative applications of cartography as is illustrated by modern car navigation systems (e.g., TomTom, Garmin, etc.), Google maps, Mashups using geo-tagging, etc. People can seamlessly zoom in and out using the interactive maps in such systems. Moreover, all kinds of information can be projected on these interactive maps (e.g., traffic jams, etc.).

Process models can be seen as the "maps" describing the operational processes of organizations. Unfortunately, *accurate and interactive process maps are typically missing* when it comes to business process management. Either there are no good maps or maps are static or outdated. Moreover, business process maps are typically *not well understood* by end users. Therefore, we propose to *automatically generate understandable business process maps using innovative process mining techniques*. By doing this, we establish a close connection between these maps and the actual behavior recorded in event logs. This will allow for high-quality process models showing what really happened. Moreover, this will also enable the *projection of dynamic information* on process maps, e.g., showing "traffic jams" in business processes.

In order to create such "business process maps", process mining techniques need to be improved. Current techniques have problems dealing with noise and infrequent behavior. Moreover, the fact that there are no negative events and that logs tend to be far from complete (i.e., contain just a sample of typical behavior) makes process discovery challenging. *Most process mining techniques do not adequately incorporate the stochastic nature of processes and their models.* Some behaviors may be frequent and easy to uncover, while other behaviors are less probable and difficult to extract from event logs.

2.2 Log-based auditing

The term *auditing* refers to the evaluation of organizations and their processes. Audits are performed to ascertain the validity and reliability of information about these organizations and associated processes. This is done *to check whether business processes are executed within certain boundaries set by managers, governments, and other stakeholders*. For example, specific rules may be enforced by law or company policies and the auditor should check whether these rules are followed or not. Violations of these rules may indicate fraud, malpractice, risks, and inefficiencies. Traditionally, auditors can only provide *reasonable assurance* that business processes are executed within the given set of boundaries. They check the operating effectiveness of controls that are designed to ensure reliable processing. When these controls are not in place, or otherwise not functioning as expected, they typically *only check samples of factual data*, often in the "paper world".

However, today detailed information about processes is being recorded in the form of event logs, audit trails, transaction logs, databases, data warehouses, etc. Therefore, it should no longer be acceptable to only check a small set of samples offline. Instead, *all events in a business process can be evaluated and this can be done while the process is still running*. The availability of log data and advanced process mining techniques enable a new form of auditing: *log-based auditing* [4]. Log-based auditing is driven by actual event data and includes automated *business process forensics*, i.e., gathering data to check whether people and organizations are operating within the boundaries set by the "de jure" models.

For log-based auditing it is also important to incorporate the stochastic nature of business processes. The fact that violations take place does not necessarily indicate a problem. Organizations need to be able to respond to unexpected events in a flexible way. Hence, deviations are a fact of life. Therefore, it is interesting to develop conformance checking techniques for stochastic models that tolerate some preset level of non-conformance.

2.3 TomTom for business processes

Navigation systems have proven to be quite useful for many drivers. People increasingly rely on the devices of TomTom, Garmin and other vendors and find it useful to *get directions* to go from A to B, know the *expected arrival time*, learn about *traffic jams* on the planned route, and be able to *view maps* that can be *customized* in various ways (zoom-in/zoom-out, show fuel stations, speed limits, etc.). However, when looking at business processes and their information systems, *such information is typically lacking*.

In existing information systems it is typically *difficult to explore the surroundings* when working on a case. In contrast, a navigation system can show the current location and possible continuations. Moreover, very few information systems support prediction. While a TomTom device is continuously showing the *expected arrival time*, users of today's information systems are left clueless

about likely outcomes of the cases they are working on. This is surprising as a lot of historic information is gathered by these systems, thus providing an excellent basis for all kinds of predictions (expected completion time, likelihood of some undesirable outcome, estimated costs, etc.). A navigation system is always *recommending a particular route*. The recommended route is not fixed and continuously changes based on the actions of the driver and contextual information (e.g. traffic jams). Note that information systems do not provide such recommendations; users are forced to work in a particular manner or get no guidance at all. As discussed in [5], process mining research should aim at supporting TomTom-like functionality.

For predictions and recommendations, stochastic models need to be discovered from event logs. These models need to be learned based on historic information and also need to incorporate the current state of the process. In principle, queuing networks and Markov models can be used. However, due to the complexity and concurrent nature of business processes, more powerful representations and analysis techniques are needed.

2.4 Business process provenance

Starting point for addressing the three challenges is the availability of high-quality event logs. We use the term *business process provenance* to refer to the systematic collection of the information needed to reconstruct what has actually happened. The term signifies that for auditing it is vital that "history cannot be rewritten or obscured". From an auditing point of view the systematic, reliable, and trustworthy recording of events is essential. Therefore, we propose to collect (whenever possible) provenance data outside of the operational information system(s) as shown in Figure 1. This means that events need to be collected and stored persistently.

3 Towards a Mature Research Discipline

After discussing some of the main challenges, we focus on process mining as a research discipline. An increasing number of process mining publications appears in journals and conference proceedings each year. Most of these publications focus on process discovery, i.e., constructing a process model (e.g., a Petri net) from some event log. From a scientific point of view this is a challenging problem and related to the classical work on the limits of inductive inference by Gold [6] and Angluin and Smith [7]. However, there are also notable differences. Unlike most of the classical work, process mining aims at higher order representations which explicitly show concurrency (e.g., Petri nets, UML ADs, EPCs, BPMN, etc.) rather than lower level representations (e.g., Markov chains, finite state machines, or regular expressions). Moreover, process mining algorithms cannot assume negative examples (i.e., there are no events stating that an activity *cannot* happen) and need to deal with issues such as incompleteness (i.e.,

if something did not happen, it may still be possible) and exceptional behavior [2].

One of the main problems when looking at the many papers on process discovery is that it is very difficult to compare the results. There are three main reasons for this.

First of all, *authors do not agree on evaluation criteria*. In fact, most process mining researchers do not evaluate their techniques in a systematic manner. For example, some authors have defined a notion of fitness, i.e., the fraction of events that can be explained by the model [8]. However, various notions of fitness exist depending on the penalties associated to problems when mapping the event log onto the model. Moreover, there are other notions of conformance, e.g., appropriateness and minimal description length techniques. These notions try to make "Occams Razor"[3] operational. For the time being, the conformance notions defined in [8] seem to be the most suitable for evaluating process mining results in a systematic manner. However, more research is needed to improve these notions and remove any representation bias.

Second, *many techniques do not incorporate the stochastic nature of processes*. Several techniques described in literature do not consider the likelihood or frequency of behavior. These techniques consider things to be possible or not. This is like showing map without distinguishing between highways and bicycle paths. Classical notions for comparing process models (e.g., trace equivalence, branching bisimulation, etc.) do not incorporate the likelihood or frequency of behavior. One of the exceptions is the approach presented in [9].

Third, there is a *lack of widely used benchmark logs*. Many authors test their techniques on self-generated synthetic logs. Mostly these are obtained through simulation. In some cases, these logs are even made by hand. From a scientific point of view, it is unacceptable to test process discovery techniques using only biased/synthetic event logs. The only way to compare process mining techniques is to test them on large collections of realistic event logs that are shared among researchers. Note that synthetic event logs are generated while having particular assumptions in mind. In reality these assumptions often do not hold. Moreover, papers presenting inferior process mining techniques get published simply because of a lack of challenging benchmark examples.

In [10] some suggestions are made to address the problems just mentioned. It is shown how a dedicated ProM plug-in can be used to compare the results of multiple process mining techniques using multiple criteria. A more recent initiative is the 3TU.Data Center (cf. `www.datacentrum.3tu.nl`). The 3TU.Data Center is a joint cooperation of the three Dutch technical universities to guarantee permanent accessibility to scientific research data in The Netherlands. One of the first initiatives will be to provide event logs for process mining. These logs can be referenced using the Digital Object Identifier (DOI). Hence, like publica-

[3] Occams Razor is a principle that states that "one should not increase, beyond what is necessary, the number of entities required to explain anything". In the context of process mining this means that we should look for the "simplest model" that can explain what is in the log.

tions, benchmark event logs get a unique ID that can be referenced and accessed easily.

Process mining is a young and exciting research discipline with high practical relevance. The *IEEE Task Force on Process Mining* aims to promote the research, development, education and understanding of process mining. This Task Force was established in the context of the Data Mining Technical Committee of the Computational Intelligence Society of the Institute of Electrical and Electronic Engineers, cf. `www.win.tue.nl/ieeetfpm/`. For readers that want to learn more about process mining, we refer to `www.processmining.org` which provides presentations, papers, videos, and software tools.

Acknowledgments. The author would like to thank all the people that contributed to the development of ProM and that promoted the use of process mining in various application domains.

References

1. W.M.P. van der Aalst, H.A. Reijers, A.J.M.M. Weijters, B.F. van Dongen, A.K. Alves de Medeiros, M. Song, and H.M.W. Verbeek. Business Process Mining: An Industrial Application. *Information Systems*, 32(5):713–732, 2007.
2. 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.
3. W.M.P. van der Aalst, B.F. van Dongen, C.W. Günther, R.S. Mans, A.K. Alves de Medeiros, A. Rozinat, V. Rubin, M. Song, H.M.W. Verbeek, and A.J.M.M. Weijters. ProM 4.0: Comprehensive Support for Real Process Analysis. In J. Kleijn and A. Yakovlev, editors, *Application and Theory of Petri Nets and Other Models of Concurrency (ICATPN 2007)*, volume 4546 of *Lecture Notes in Computer Science*, pages 484–494. Springer-Verlag, Berlin, 2007.
4. W.M.P. van der Aalst, K.M. van Hee, J.M. van der Werf, and M. Verdonk. Auditing 2.0: Using Process Mining to Support Tomorrow's Auditor. *IEEE Computer*, 43(3):102–105, 2010.
5. W.M.P. van der Aalst. TomTom for Business Process Management (TomTom4BPM). In P. van Eck, J. Gordijn, , and R. Wieringa, editors, *Advanced Information Systems Engineering, Proceedings of the 21st International Conference on Advanced Information Systems Engineering (CAiSE'09)*, volume 5565 of *Lecture Notes in Computer Science*, pages 2–5. Springer-Verlag, Berlin, 2009.
6. E.M. Gold. Language Identfication in the Limit. *Information and Control*, 10(5):447–474, 1967.
7. D. Angluin and C.H. Smith. Inductive Inference: Theory and Methods. *Computing Surveys*, 15(3):237–269, 1983.
8. A. Rozinat and W.M.P. van der Aalst. Conformance Checking of Processes Based on Monitoring Real Behavior. *Information Systems*, 33(1):64–95, 2008.
9. A.K. Alves de Medeiros, W.M.P. van der Aalst, and A.J.M.M. Weijters. Quantifying Process Equivalence Based on Observed Behavior. *Data and Knowledge Engineering*, 64(1):55–74, 2008.

10. A. Rozinat, A.K. Alves de Medeiros, C.W. Günther, A.J.M.M. Weijters, and W.M.P. van der Aalst. The Need for a Process Mining Evaluation Framework in Research and Practice. In A. ter Hofstede, B. Benatallah, and H.Y. Paik, editors, *BPM 2007 International Workshops (BPI, BPD, CBP, ProHealth, RefMod, Semantics4ws)*, volume 4928 of *Lecture Notes in Computer Science*, pages 84–89. Springer-Verlag, Berlin, 2008.