A Framework for Applying Process Mining Techniques in Software Process Assessments

Arthur Valle, Eduardo Portela, Eduardo Rocha Loures, José Marcelo Cestari
Pontifical Catholic University of Parana
Brazil

Abstract

Process assessment methods typically use the following approaches to collect information about processes in an organization: questionnaires, document review, interviews and demonstrations. It is possible to note that existing processes assessment methods, for example SCAMPI (Standard CMMI Appraisal Method for Process Improvement), do not use contemporary data analysis techniques such as processes mining, text mining or data mining. Process mining is one of the techniques that can be used to assist process assessments, in order to conduct assessments with greater coverage and reliability while keeping lower cost and effort. The aim of this paper is to present a framework to apply process mining techniques in software process assessments. A systematic literature review was conducted to identify papers describing process mining applied to software process assessments. Case studies are planned to be conducted in order to validate the use of the framework.

Keywords
Process mining, Software process assessment, Data collection and analysis.

1. Introduction

Process assessment is the examination of one or more processes by a trained team using an assessment model as the basis for determining, at least, strengths and weaknesses [1]. In the context of the software industry, SCAMPI-Standard CMMI Appraisal-Method for Process Improvement is an example of process assessment method [2].

One aspect that all process assessments have in common is that they are often very laborious and expensive. Studies such as Rout et al (2007) point out that the most costly activity during a software process assessment is data collection, representing 47% of the total effort [3]. Moreover, this fact is aggravated since that in process assessments just data samples are normally checked, i.e. this percentage would be even greater if larger samples were considered.

To prepare data required for a process assessment, many organizations invest a significant amount of time, effort and resources. In fact, one gets the impression that some organizations spend more resources on data preparation than in support of all other assessment activities [2]. Therefore, the overall result is usually significant expenses in terms of time, effort and financial resources to support data collection activities for an assessment.

However, these days, detailed information about processes is being recorded in the form of event logs, transaction logs, databases, etc. In this sense, in a process assessment, is no longer justifiable check only a small set of processes. Instead, all events in a process can be evaluated. The presence of event logs and process mining techniques enable an additional approach to existing techniques in process assessment methods: instead of a small sample of instances, the entire process and all its instances can be considered, as long as this represents low costs, naturally.
Besides the cost aspect, there is another question: considering data collection activities in existing assessment methods, the following techniques are used for gathering information about the running processes in an organization: questionnaires, document review, interviews and demonstrations. However, no contemporary data collection and analysis techniques such as data mining, text mining or process mining are used.

2. Objectives
The application of newer techniques of collecting and analyzing data would allow assessments with more objectivity, accuracy, depth and coverage of aspects related to the execution of processes, such as duration and sequence of activities, start and end dates and records of who were the real performers.

Therefore, this research project proposes the integration of process mining techniques on current software process assessment methods. In this approach, a greater emphasis will be placed in the event log of software processes. Based on these event logs, it is possible to understand the story in a complete, economical, reliable and accurate manner, thereby contributing to the collection and analysis of data, critical activities in any software process assessment method.

Conformance checking - a type of process mining technique - identifies deviations, for example, highlighting parts of the “as-is” model where compliance is low or pointing situations that deviate from “to-be” models or business rules. This is highly relevant for the purposes of process assessment, since it can be used to verify which rules are violated and where and when people did not perform the process as specified.

Additionally, in the literature, there are few cases of application of such techniques in software process assessment methods.

The main objective of this paper is to present an overview of a framework for the application of process mining techniques in software process assessment methods. This framework aims to enable greater coverage, efficiency and depth of analysis than traditional techniques of collecting and analyzing data, while maintaining similar levels of effort, cost and time.

The remainder of this paper is organized as follows: the next section provides background on software process assessment and process mining techniques. The framework is then described in its two views, dynamic and static. Finally, conclusions and future research are discussed.

3. Background
In this section, characterization of data collection and analysis activities in software process assessments, process mining, and integration of both themes are presented.

3.1 Data Collection and Analysis in Software Process Assessments
In the literature, there are several definitions for the term process assessment. For example, Rout et al (2007) define process assessment as the disciplined examination of the processes used by an organization against a set of criteria to determine the ability of these processes performing within the quality, cost and schedule objectives [3].

A process assessment method includes the activities that must be performed in order to conduct a process assessment. An example of process assessment method is SCAMPI - Standard CMMI Appraisal Method for Process Improvement [2], which is used to identify strengths, weaknesses and ratings relative to CMMI reference models [1].

In SCAMPI method, data collection and analysis is characterized as shown in figure-01 "Data collection and analysis in SCAMPI method":
First, SCAMPI method emphasizes the concept of data collection strategy: "A well-defined data collection strategy is important for assessment planning, since it provides the basis for detailed planning of data collection and examination of objective evidence" [2].

The data collection strategy, which seeks an objective characterization of the process used by the organization, impacts assessment on aspects such as a) amount of time and effort expended by the organization in assessment preparation, b) team's ability to make accurate judgments, c) usability and accuracy of the assessment results and d) total cost of onsite phase. In SCAMPI method, data collection strategy defines the high level structure for data collection, including: a) selection of data collection approach, b) when data will be collected, c) what techniques will be employed to collect data, d) responsibilities for data collection.

Beyond the concept of strategy, SCAMPI method defines approaches to data collection, as shown in figure 01. Currently, there are three approaches:

- **Verification-based**
  - An assessment that the focus of the team is to check the set of objective evidence provided by the organization being assessed before the onsite period, in order to reduce the amount of evidence objective search and discovery during onsite phase [3].

- **Discovery-based**
  - An assessment in which a limited number of objective evidence is provided before the onsite phase by the organization being assessed, and during the onsite phase, assessment team investigates and discovers most of the evidence objective necessary for achieving sufficient coverage of the reference model practices [3].

- **Managed Discovery**
  - A phased approach to data collection, starting with a request for a pre-determined set of initial data (artifacts), followed by a set of iterative requests based on the evaluation by the team, of remaining work products and gaps in evidences [3].

According to figure 01, another major element is the data collection plan, which is based on the defined data collection strategy. A data collection plan helps to document and communicate the general definitions of data collection for the assessment.

In addition, process assessments depend on an aggregation of information that is collected via defined types of objective evidence. Objective evidences are artifacts or affirmations used as indicators of implementation or institutionalization of practices in a model [2]. The assessment team notes, hear or read information that is transformed in notes, and then on characterizations of compliance or gaps in implementation of practices and later on preliminary results. These results are validated by organizational unit before becoming final findings. In general, in a process assessment, there are the following types of objective evidence:
Artifacts: tangible objective evidence, indicative of work being performed, which is the primary output of a model practice or a consequence of the implementation of a model practice [2]. Artifacts can include organizational policies, minutes of meetings, records, reports, review results or other work products.

Affirmations: oral or written statements, provided by executors of the practice to confirm or support the implementation (or lack of implementation) of a model practice [2]. Affirmations are usually collected through a variety of techniques:
- Interviews
  - An interview is a meeting of the members of the assessment team with the participants of the assessment for the purpose of collecting information on existing work processes [2].
- Presentations
  - In a process assessment, a presentation is a source of objective evidence, which includes information prepared by the organization and delivered either visually or verbally to the assessment team to assist in the understanding of organizational processes and practical implementation of the reference model [2].
- Demonstrations
  - Technique of data collection where there is an interactive demo (or walkthrough) of a tool by the person that uses such tool as part of the execution of his/her process.
- Questionnaires
  - Questionnaires are a special type of instruments. Instruments are provided by the organizational unit to inform assessment team about processes implemented in the organization and how they relate to the reference model.

3.2 Process Mining
Another key topic of this research is Process Mining. The idea of process mining is to discover, verify and improve real processes by extracting knowledge from event logs that are readily available in today’s information systems [4]. Process mining can answer questions such as "What does the process looks like?” or “How much does actual process deviate from defined process?". The assumption is that during a process assessment, these same questions are asked and should be answered, and therefore process mining can be recognized as a valuable way to collect and analyze objective evidence required for software process assessments [5].

According to figure 02 above, extracted from Process Mining Manifesto [4], process mining allows automatic derivation, called process discovery, of the “as is” process model, that is as well as conformance checking, and further process enhancement [6]. These are the three main types of process mining, as shown in figure 03 - "process mining types", also extracted from Process Mining Manifesto:
Figure 03 - Process mining types

- Process Discovery: From an event log, a “as is” process model is identified.
- Conformance Checking: An existing process model is compared to an event log of the same process. Compliance checking can be used to verify if the process as described in the records of events is consistent with the models and vice versa (see figure 02) [4].
- Enhancement (or extension): A process model is improved using information extracted from a log [4].

3.3 Integration of Process Mining and Software Process Assessment

As part of the literature review of this research, a systematic literature review was conducted, according to the guidelines proposed in [7], in order to investigate the existing integration between topics a) software processes assessment and b) process mining. According to Kitchenham and Charters, a systematic literature review is a way to evaluate and interpret all relevant research on a particular research question, area or phenomenon of interest [7].

The following criteria and procedures for selection of studies were used: “Only will be selected peer-reviewed publications in journals or conference proceedings, in English, that demonstrated proposed or actual application of process mining in software process assessment methods”. "Process assessment" AND "process mining" terms were searched in the following databases: ScienceDirect (sciencedirect.com); Citeseer library (citeseer.ist.psu.edu); IEEE Xplore (ieeexplore.ieee.org); Wiley Online Library (onlinelibrary.wiley.com); Scopus (scopus.com); ACM Digital Library (dl.acm.org).

Since first tests showed that no work would be selected, it was necessary to expand the selection criteria for the following: “Only will be selected peer-reviewed publications in journals or conference proceedings, in English, that demonstrated application of processes mining in software processes”. Even that, only 06 of the 26 resulting papers were selected.

None of the selected papers explained which process assessment methods were used. Due to the unfavorable results found, an additional search on Google Scholar was conducted using the same terms. Therefore, some more relevant papers to both research studies were found:

- The first one is the thesis of Samalikova, entitled Process Mining Application in Software Process Assessment [5]. The thesis objective was to promote the use of process mining in software process assessment and improvement. Her conclusion was that techniques for collecting information derived from process mining can be applied to improve the data collection on software process assessment [5]. However, conclusion was reached by qualitative judgment without objective criteria. In addition, process mining techniques that should be used were not nominally listed.
- Another paper found was the dissertation of Cruañas, entitled Process Mining Opportunities for CMMI Assessments [8]. The objective was to investigate the literature concerning support tools of key process areas of CMMI to find out if it is possible to use process mining to improve the assessment of CMMI key process areas and, by extension, the assessment of CMMI in general. His conclusion was that process mining can not only help improve the current CMMI assessments, but can also be a useful tool to assist data collection [8]. However, conclusions were based on the generalization of processes mining techniques
and perspectives and without using any objective criteria. Moreover, no process mining technique in particular was pointed out.

- A third paper found is from Rubin et al (2007), who developed a framework for process mining for software processes [9]. Different aspects of processes mining were addressed, such as control perspective (that captures the order in which activities are performed), information perspective (which captures data, documents and information required or produced by an activity) and organizational perspective (which identifies people or roles that perform a particular activity). Some algorithms that can be applied were presented in the paper, as well as in Samalik’s paper. On the other hand, authors also reported a lack of algorithms to generate formal models for software processes [9].

It can be observed that in software development, records obtained from support tools, such as configuration management system and development tools can be used to generate the actual process model being practiced. This model can be analyzed, verified and subsequently optimized and managed. In other words, process mining can be used not only to discover but also to monitor and improve software processes using real data repositories. However, use of process mining techniques on software development processes assessment methods is fairly crude which clearly justifies the development of a framework that guides application of process mining techniques in process assessment methods.

4. Framework

4.1 Description
The framework described in this paper, is a structure intended to serve as a guide for applying process mining techniques in software process assessments. The framework is represented by two views: dynamic view and static view.

4.2 Dynamic View of the Framework
Dynamic view represents how the process flows. Figure 04—“dynamic view of the framework”, adapted from [10], considers M1 as the selected reference model, for instance, CMMI-DEV (Capability Maturity Model Integration for Development) [1]. M1 is used as a basis to define “to-be” process model M2. M2 is executed by the organization and teams, generating record files, log files, register of performers, etc. Subsequently, process mining techniques, especially alpha-algorithm, are applied in these type of data, in order to discover “as-is” process model M3.

![Dynamic view of the framework](image)

Figure 04 – Dynamic view of the framework, adapted from [10]

In the dynamic view the compliance (or conformance) aspect is also represented, as shown by double arrows in figure 04. Compliance 1 refers to compliance between M2 and M1, which reflects the degree that M2 conforms to M1 and vice-versa. For example, one can verify if each best practices of a reference model (M1) has a correspondent implementation in the defined process (M2). Typically it involves business rules and business rules checking. Compliance 2 refers to compliance between M3 and M2, i.e. the degree that M3 conforms to M2 and vice-versa. For example, one can verify if “real” process (M3 deviates from the defined process model (M2) or how distance is the defined process (M2) in relation to its real execution (M3). Typically it involves the processing of event logs by
process mining tools and algorithms. Compliance 3 refers to compliance between M3 and M1, which means the degree that M3 conforms to M1 and vice-versa. For example, one can verify if all best practices of a reference model (M1) are being followed in the real execution of process (M3). Typically it also involves business rules and business rules checking.

4.3 Static View of the Framework

The static view of the framework is shown in figure 05—“Static view of the framework”. It consists of elements and functions, represented respectively by cylinders and rounded rectangles. Static view covers aspects from process assessment, presented in white color icons and aspects from process mining discipline, presented in blue color icons in fig. 05. Process assessment aspects (i.e elements and functions) come from process assessment methods and are aligned with data collection and analysis characteristics as described in section 3.1. Process mining aspects come from process mining area, as described in section 3.2.

Figure 05 – Static view of the framework

4.3.1 Elements of Static View of the Framework

In this section, elements of static view are described.

- Reference model (or best practices model)
  - Reference models, also known as best practices models, are collections of best practices that help organizations improve their processes [1]. As stated early, CMMI is an example of a reference model. In the context of this framework, data resulting of application of process mining techniques is verified to check compliance with best practices described in these reference models.

- Raw data
  - Raw data generated by the execution of processes in an organization. Usually, they are manual records, documents/artifacts in a repository or records generated by the use of software tools. Raw data are filtered and transformed into objective evidence.

- Objective evidence
  - Objectives evidences are artifacts or affirmations used as indicators of implementation or institutionalization of the practices of a reference model [2]. In special, artifacts are filtered and transformed into event logs.

- Event log
An event log is a set of data that is derived from artifacts and contains relevant data to process mining algorithms, such as register of instances, activities, timestamp and performers.

- **De jure** process model
  - *De jure* model are defined process models that describe how a business process should be executed, also called “as is” process model.

- **De facto** process model
  - *De facto* model are process models, obtained through process discovery technique, based on data from the event log and thus describe the actual observed behavior of a process, also called “to be” process model.

- **Findings**
  - Strengths and weakness statements that are inferences drawn from corroborated objective evidence. As the main outcomes of a process assessment, they represent the current situation of the appraised organization.

### 4.3.2 Functions of Static View of the Framework

In this section, functions of static view are described.

- **Data Collection**
  - Raw data is collected and filtered in order to become objective evidence. Typically, it is done manually by a reference model expert.

- **Examine, document and verify objective evidence**
  - Objective evidence are examined, documented, verified and derived into assessment findings.

- **Data Filtering**
  - It is the conversion of artifact-type data in a filtered event log that can be read by process mining tools. It is done through process mining conversion tools.

- **(Process) discovery**
  - It is the derivation of a de facto process model from an event log. It is done by process mining tools, such as ProM [11].

- **Conformance checking**
  - It is the conformance check used to verify if two elements conforms each other. In the context of this framework, conformance checking function could be subdivided into:
    - **Model to model conformance checking**
      - It is the conformance check used to verify if the process model derived from event log (i.e. *de facto model*) is in accordance with the discovered process model (i.e. *de jure model*) and vice versa.
    - **Event log to model conformance checking**
      - It is the conformance check used to verify if event logs are in accordance with de jure model and vice-versa. This comparison shows where actual instances deviate from the modeled process.

- **Business rules checking**
  - It is the conformance check used to verify if process mining resulting data (i.e event log, de jure model and de facto model) is in accordance with business rules extracted from the reference model, and vice-versa. Business rules define how a business process must be executed. In the context of this framework, conformance checking function could be subdivided into:
    - **De jure model to rules conformance checking**
      - It is the conformance check used to verify if de jure process model is in accordance with the business rules, and vice-versa.
    - **De facto model to rules conformance checking**
      - It is the conformance check used to verify if de facto process model is in accordance with the business rules, and vice-versa.
    - **Event log to rules conformance checking**
      - It is the conformance check used to verify if event logs are in accordance with the business rules, and vice-versa.

- **Analysis**
  - It is the analysis of process mining resulting data in order to contribute to drawing assessment findings.
5. Conclusion
Process assessment methods are built on traditional approaches of data collection and analysis, such as interviews and demonstrations of tools. More effective techniques, such as process mining techniques, are demanded in order to transform existing process assessment methods more productive and economically viable. In the other hand, research on process mining area is growing faster, and new algorithms are being developed regularly. Even assessors and process analysts who already use process mining techniques may find trouble in identifying which process mining algorithms to use and how to apply them to answer the typical “questions” of a software process assessment. In this sense, a well-structured framework that guides application of process mining techniques in software process assessments is clearly needed.

6. Future Research
The framework presented in this paper clearly needs to be refined and tested on a real process assessment event. For example, multiple case studies that will test the use of the proposed framework in SCAMPI class A appraisal are being envisioned. Although those two actions are being addressed, they are out of the scope of this paper.

References
11. van der Aalst, W. M. et al., 2007, “ProM 4.0: comprehensive support for real process analysis”. In Petri Nets and Other Models of Concurrency—ICATPN 2007 (pp. 484-494). Springer Berlin Heidelberg.