Applying Process Mining in SOA Environments

Ateeq Khan, Azeem Lodhi, Veit Köppen, Gamal Kassem, and Gunter Saake

School of Computer Science, University of Magdeburg, Germany
ateeq,azeem,veit.koeppen,gamal.kassem,saake@iti.cs.uni-magdeburg.de

Abstract. Process mining is an emerging analysis technique, which extracts process knowledge from data and provides various benefits to organizations. In Service Oriented Computing environment, different services collaborate with others to carry out the operations and therefore overall picture of operations and execution is not clear. Process mining extracts the information from log files of systems, as recorded during executions, and depicts the reality. In order to apply process mining, extraction of process trace data from log files is a pre-requisite step. A case study demonstrates the practical applicability of our proposed framework for extraction of the process trace data from application systems and integration portals.

Keywords: Business process analysis, Process trace data, Log files, SAP Process Integration, Process mining.

1 Introduction

In Service Oriented Computing, different services collaborate with one another to perform tasks. The same situation occurs in enterprises where business operations are carried out by different service interactions. This is due to several characteristics of involved elements, availability of services, resources, employee's experiences. Therefore, the operations or business processes can be executed in different ways as compared to a pre-defined way. This deviation motivates to apply process mining for business process analysis.

Business process analysis is a basic step for adaption and improvement of systems. Adaptive systems require an actual and consistent picture of the current environment, Process mining is a good candidate for this purpose [1]. Similarly, in case of embedded systems, process mining can be applied for monitoring interactions of devices, performance analysis and execution of tasks. In literature, benefits of process mining with respect to different perspectives are highlighted, e.g. process discovery [2], conformance checking [3] (comparison between designed model and the actual model in execution), social network analysis (who is responsible and collaborating with whom) [4]. These analyses help to improve the future execution of processes and thus to achieve the business goals in an efficient way. To apply process mining, the extraction of process data from log files is a basic and challenging step [5,6]. We address it and provide a framework in Sect. 3 to extract process trace data. Our proposed framework is illustrated with a case study in Sect. 4 SOA-based SAP Netweaver Platform.

A. Dan, F. Gittler, and F. Toumani (Eds.): ICSOC/ServiceWave 2009, LNCS 6275, pp. 293–302, 2010.

2 Basics

In this section, we give an overview of the basics required for this paper. First, we introduce process mining. Furthermore, we give some basic insights to the SAP environment to understand the execution environment of our case study.

2.1 Process Mining

Process mining aims to extract process knowledge and to discover process models from executions of business processes as recorded by information systems [5]. The overall approach of process mining is shown in Fig. 1.

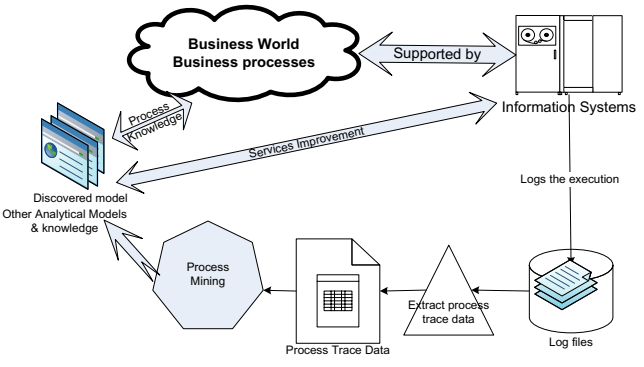

Fig. 1. Overall Approach of Process Mining

Log Files. Information systems execute business processes and record the data about elements in the form of log files or as records in database tables. In this work, we refer to log or trace files as interchangeable terms. Log files contain data about occurrences of events, provided inputs, processing information, generated outputs, message exchanges, usage, and condition of resources during execution. These log files are starting point for process mining.

Process Trace Data. To apply process mining, it is assumed that the information system records data of events, cases, time stamp of an event, sequence of activities, performer, and originator [4]. These log files also contain unstructured and irrelevant data, e.g. information on hardware components, errors and recovery information, and system internal temporary variables. Therefore, extraction of data from log files is a non-trivial task and a necessary pre-processing step for process mining. Business processes and their executions related data are extracted from these log files. Such data are called process trace data. For example typical process trace data would include process instance id, activity name, activity originator, time stamps, and data about involved elements.

Conversion, Mining, and Analysis. Extracted data are converted into the required format, depending on the process mining tools or algorithms. For social

network analysis, information on originator or performer of activity is important, performance based analysis requires for instance time related data. Therefore, the perspective of analysis depends on available data. Several process mining tools and techniques for process knowledge discovery are discussed in [7].

2.2 SAP Netweaver Platform and SAP Process Integration

SAP Enterprise resource planning (ERP) systems are used to provide integrated data and processes across all departments. SAP NetWeaver provides a facility for integrating SAP and non SAP systems. SAP NetWeaver uses open integration (providing interoperability) to connect with other systems or existing solution.

SAP Process Integration (SAP PI) is an Enterprise Application Integration product. It resides in process integration layer of SAP NetWeaver to integrate internal and external processes. SAP PI system works as a central hub and all systems communicate with each other through it. It removes inconsistencies between heterogeneous systems. These inconsistencies arise due to different requirements of formats, interfaces, protocols, and connectivity between SAP and non-SAP applications (for A2A and B2B systems).

3 Framework for Extracting Process Trace Data

We depicted a framework in Fig. 2 for extracting process trace data from SAP NetWeaver log files. However, this is not only applicable to a SAP system, but it can also be utilized in other platforms for process mining. The integration portal maintains the logs of different system interactions. Firstly, logs are collected from information systems and integration portals. Either the logs are collected manually from each involved system or the integration portal provides the services for log collection.

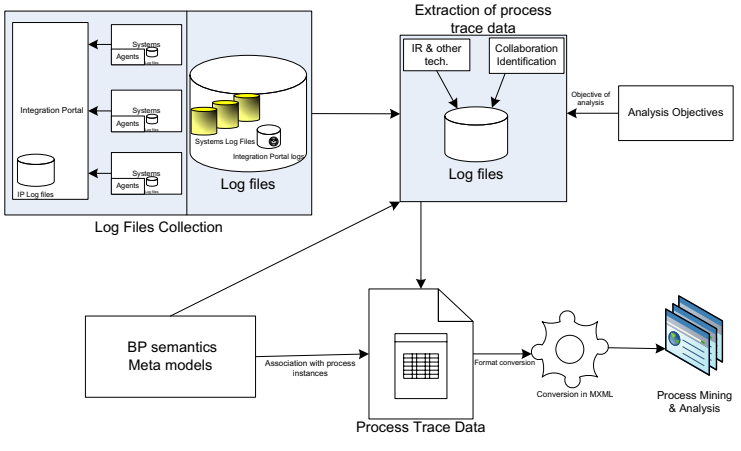

Fig. 2. Extracting process trace data in heterogeneous environment

Different system vendors record and describe the events according to their own standards, internal structures, and languages [8]. Some vendors provide meta models to describe the structure of their log files. These descriptions are used for understanding and extraction of the process trace data from log files. Some meta models of workflow management systems are discussed in [7].

Sometimes, meta models are not available or logging notation (standards, meta models) is not followed completely. This requires the understanding of systems and business semantics. In such situation, logs are manually analyzed to extract the structure. For example data related with events are confounded with system data, and spanned across various log files and tables.

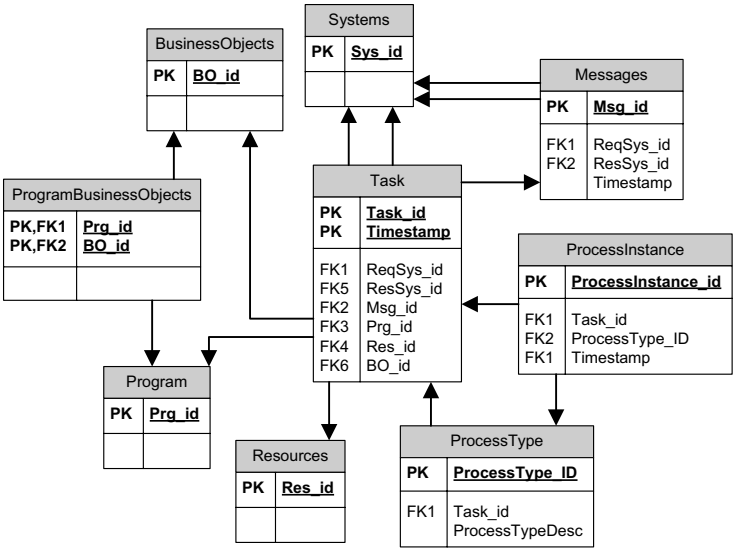

Fig. 3. Generic Meta Model of log files and process trace data

A generic meta model is presented to extract and relate the data of log files in Fig. 3. In the meta model, the entity *ProcessType* describes the business process like receiving an order, manufacturing, and shipping. *Task* is the atomic unit of work, and can belong to more than one *ProcessType* entity (e.g. getCustomerDetails can be used in receiving order as well as in shipping business process). *Process instance* is the representation of a customer's request. The request passes through different phases of a process for operations, and therefore, it contains more than one *ProcessType* and *Task*. *Resources* perform different tasks on the process instance and modify/change the state of *Business Objects* attached to them. The *Programs* record the change of state of business objects.

In heterogeneous environments, different systems are involved in the business process execution. These systems communicate with each other to execute business processes within or outside the enterprise. The content of a message describes the nature of operations and involved objects. Integration Portal log files

contain data such as requesting systems, contents of request, response, and performer, time stamps of requests/responses, message ids, and message contents, and resources. Such data can be used as process trace data to identify the collaboration between systems. Organizational business data (operational) is also used for relating/associating the data. For example in customer-order scenario, customer order number can be used to relate the customer name, ordered items, and other data. Data from log file are extracted and used to build the execution instances.

Various elements are not considered, e.g. activities, screens (and GUIs), states of processes, and business objects. We present a quite generic model. It is extendable based on analysis objectives, for application usage mining, the *Program* entity is extended and related with the GUIs presented to users, elements contained, and user actions.

4 Case Study and Extraction of Process Trace Data

We explained our framework in a case study. Different SAP systems and web services are integrated by SAP PI to complete the business processes.

The case study is a simple supply chain scenario; a customer visits the website and places an order. Order items received as a file in SAP PI from the website through http-file adapter. This file is transferred, using mail adapter, from SAP PI as formatted email to the distributor. Stock is checked and if the item is not available then a production order, containing item details, is issued from SAP R/3 system using RFC and IDOC adapter to SAP IDES system used at the manufacturing site. To ship, the total weight and customer destination are forwarded to shipping web service, which calculates a freight charge for the order and notifies the customer after shipping.

We use SAP PI for analysis and monitoring of systems (services). User trace loggings are activated in SAP systems and monitoring is done during execution. After execution, tracing is deactivated, and we collect trace files.

We develop a tool(output is depicted in Fig. 5) to extract process trace data from SAP trace files. Development of the tool consists of following phases and steps in 5. We divide the process into four phases:

4.1 Contents and Structure Analysis in Log Files

SAP does not provide meta models of log files, so we study log files and different repositories of SAP systems for extraction. In normal cases, SAP transaction data are stored in log files and used for process mining. Analysis of log files enables us to find patterns of characters that occur with events, and these characters help in extraction of process trace data from the systems. For process trace data, diag processor (Dialog work processes deal with requests from an active user to execute Dialog steps) and task handler component are required. Task handler contains user information and time of activities. Detailed analysis of SAP trace files and important structure used for extracting required data from trace file are described in [9].

Screen Messages: Screen messages are shown to users during business process execution. These messages indicate flow of the system. Information stored with these messages are message type, message time, process component, associated user, user terminal, server, and executed transactions.

Extraction of Database Fields: Elements in trace file are represented as entity types, also called etypes. Etypes have data fields attached with them, which belong to screen elements, e.g. text fields or data displayed from the database. Etypes have a specific structure and database field names are extracted after meeting conditions, e.g. a mark 'X' present on line relative to etypes type line in trace file. The database fields are assigned descriptive names. Thus, the semantics and structure of log files detected in this part help to identify important elements in log files for process trace data.

4.2 Structure Analysis in Integration Portal

We identify that trace files from SAP PI do not contain information on executed transactions or any exchanged data, e.g. when RFC module or IDOC adapter data are exchanged. In SAP PI, the logs are maintained but at the user interface level only very limited view is provided. Furthermore, trace data, of exchanged messages between systems, are stored in SAP PI system tables. In Fig. 4, a meta model for extraction of data is presented. For each message exchange, communication systems are involved as message sender or receiver. Message's contents are mapped between communication systems by mapping relations and condition involved for these mappings. Each message has a message id, associated sender, receiver, and message payload. Payload is transferred data and contains attributes, which we use for process mining and different perspectives of analyses. The time stamps of exchanged messages and other data provide connections where processes are collaborating with the system.

4.3 Collaboration Data from Integration Portal

Database tables and descriptions identified in previous steps are used to extract the data. With SAP Developer role in SAP PI, ABAP programs can be written to get data from these tables and store them in separate files. This will make the extraction of data for evaluation easier. Depending on the requirements, tables are further investigated and necessary data are extracted.

4.4 Extraction of Data from Log Files

In this step, data are extracted from log files. Data may consist of user details, screens displayed to a user, transaction executed, confirmation or error messages, and data objects against database fields found in section 4.1. These elements are extracted and saved sequentially. Extraction of elements follows a specific structure in the trace file. Extracted data may contain redundant information, which we remove by further processing.

Extracted process trace data are used to generate case instances and to use in process mining tools as explained in Section 5.

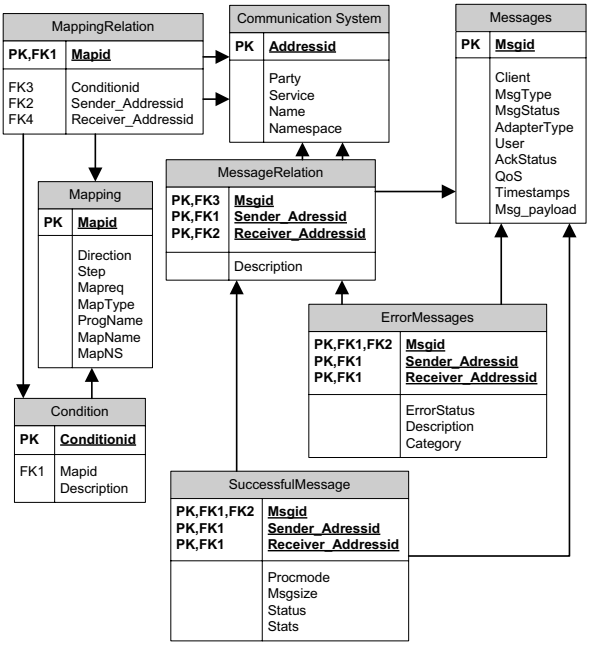

Fig. 4. SAP PI meta model for processed trace data

5 Relationship between Data and Process Mining

A primary key attribute is selected from the process trace data, e.g. key attribute may be order id, customer no., application no., or product id. All extracted information is sorted on this key attribute and written as cases. Therefore, the extracted data are correlated based on these and case instances are built. Business semantics and meta models presented in earlier sections can be used for this purpose. The extracted information is converted into a suitable format for mining and analysis tools. We provide the extracted data into two formats. In one format, cases are assigned with specific functions of log files, while in other format case numbers, object values, and functions, in which they are executed are described, see Fig. 5. We add SAP trace data conversion plug-in in ProM Import Framework [11] and use it conversion of our data into Mining XML (MXML) format. ProM[1] tool is used for analysis and mining. The actual process model is discovered. Examples of two different business scenario models are depicted in Fig. 6. Other analyses can also be done as discussed in Section 2.1. This provides the opportunity to analyze the business processes in different perspectives and directions for improvement.

[1] www.processmining.org

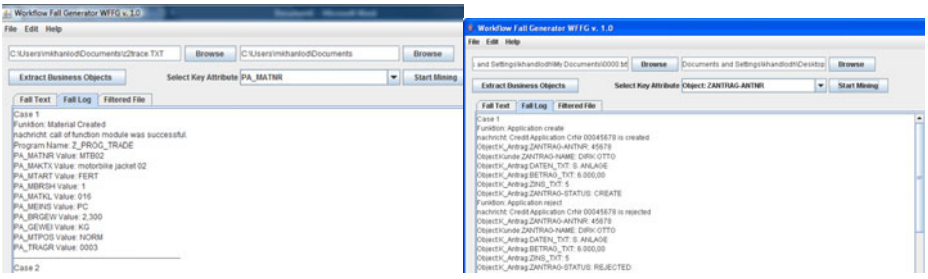

Fig. 5. Case data

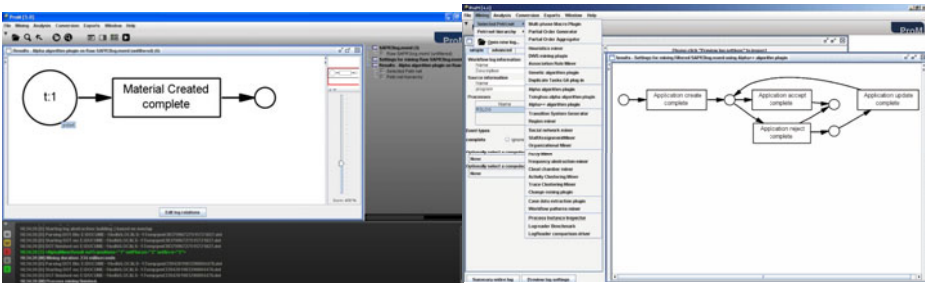

Fig. 6. AS-IS models from ProM tool [10]

6 Related Work

Process mining is an active research area in business process analysis and discussed in many papers [5,7,12,13]. Most of the papers focus on different kinds of analyses and benefits that can be achieved. These approaches assume that process trace data are already available. Extracting process trace data is a challenge and important as discussed in [5,6]. Different approaches are proposed to address this challenge. We categorize them in three major approaches:

Table-based approaches. Attempts are made to extract process data from ERP systems log files, e.g. SAP, PeopleSoft. In [14], the author tries to extract process trace data from SAP R/3 log files, but it was not fully satisfying because process execution information is too detailed due to the usage at the transaction level and scattered in many tables. In [15], authors built a tool for constructing process instances from business process executions in SAP systems. They use SQL statements to identify used business objects in SAP tables and correlate them with events to construct instances. However, this table-based approach provides less data for process mining analysis as not all data are stored with business objects in tables.

Data warehouse based approaches. Data warehouse technique is introduced in [16]. HP Process Manager (HPPM) system logs are used, so this approach is not generic. The work is further extended as architecture for analysis and monitoring in [17]. The architecture is related to our presented framework, but we focus more towards the extraction of process trace data from log files rather than analysis techniques. Information from process logs are extracted by querying log tables and stored in a process data warehouse. Mining algorithms are applied in the data warehouse.

Log files based approaches. In [7,8], meta models are proposed to correlate elements of log files. The difference from our approach is that they concentrate only to extract a small part of process trace data from log files, e.g. case id, task, timestamps, focus only for process discovery, and do not address the challenge in heterogeneous environments. We present more generic work and deals with log files as well as log tables in heterogeneous systems. In [18], data from log files are used to analyze the interaction behaviour of users with applications.

7 Conclusion and Outlook

We proposed a framework to apply process mining in heterogeneous environments. We develop a meta model to extract more process trace data from logs, so process mining can be applied to those systems, which are not process-driven. SOA based platform (SAP NetWeaver) is used as a case study to show the applicability of our framework and process mining in a SOA environment. Monitoring components of SAP NetWeaver are discussed for the extraction of process trace data. We also developed a tool for the extraction of process trace data from the logs of involved systems. We believe that this work provides a guideline to apply process mining in different SOA based systems and steps toward better analyses and monitoring services.

It would be interesting to investigate cases in which more than one integration portal is involved. Organizations may not be interested to share their data. Privacy preserving methods should be developed for this purpose.

Acknowledgment. Ateeq Khan is supported by scholarship from federal state of Saxony-Anhalt, Germany. Veit Köppen is funded by the German Ministry of Education and Science (BMBF), project 01IM08003C. The presented work is part of the ViERforES[2] project.

References

1. van der Aalst, W.M.P., Günther, C., Recker, J., Reichert, M.: Using process mining to analyze and improve process flexibility. In: Latour, T., Petit, M. (eds.) Proceedings of the CAiSE 2006 Workshops / 7th Int'l Workshop on BPMDS 2006, Namur, June 2006, pp. 168–177. Presses Universitaires de Namur (2006)

[2] www.vierfores.de

2. Weijters, A.J.M.M., Maruster, L.: Workflow mining: Discovering process models from event logs. IEEE Transactions on KDE 16, 2004 (2004)
3. van der Aalst, W.M.P.: Business alignment: using process mining as a tool for delta analysis and conformance testing. Requir. Eng. 10(3), 198–211 (2005)
4. van der Aalst, W.M.P., Reijers, H.A., Song, M.: Discovering social networks from event logs. Comput. Supported Coop. Work 14(6), 549–593 (2005)
5. van der Aalst, W.M.P., Weijters, A.J.M.M.: Process mining: A research agenda. Computers in Industry 53, 231–244 (2004)
6. van der Aalst, W.M.P.: Challenges in business process analysis. In: Filipe, J., Cordeiro, J., Cardoso, J. (eds.) Proceedings of the 9th ICEIS. Lecture Notes in Business Information Processing, vol. 12, pp. 27–42. Springer, Heidelberg (2007)
7. Muehlen, M.Z.: Workflow-based Process Controlling. Foundation, Design, and Implementation of Workflow-driven Process Information Systems. Advances in Information Systems and Management Science, vol. 6. Logos, Berlin (2004)
8. van Dongen, B.F., van der Aalst, W.M.P.: A meta model for process mining data. In: Conference on Advanced Information Systems Engineering (CAiSE) Workshops, vol. 161, p. 209 (2005)
9. Khan, A., Lodhi, A.: Analysis of Business Processes in Heterogeneous Environment: SAP as a Use Case. Master thesis, School of Computer Science, University of Magdeburg (February 2009)
10. Dongen, B., Medeiros, A., Verbeek, H.M.W., Weijters, A.J.M.M., van der Aalst, W.M.P.: The proM framework: A new era in process mining tool support. In: Ciardo, G., Darondeau, P. (eds.) ICATPN 2005. LNCS, vol. 3536, pp. 444–454. Springer, Heidelberg (2005)
11. Günther, C., van der Aalst, W.M.P.: A generic import framework for process event logs. In: Eder, J., Dustdar, S. (eds.) BPM Workshops 2006. LNCS, vol. 4103, pp. 81–92. Springer, Heidelberg (2006)
12. Greco, G., Guzzo, A., Manco, G.: Mining and reasoning on workflows. IEEE Transactions on KDE 17(4), 519–534 (2005)
13. Tiwari, A., Turner, C., Majeed, B.: A review of business process mining: State of the art and future trends. BPM Journal 14, 5–22 (2008)
14. van Giessel, M.: Process mining in sap r/3. Master's thesis, Eindhoven University of Technology, Eindhoven (2004)
15. Ingvaldsen, J., Gulla, J.: Preprocessing support for large scale process mining of sap transactions. In: ter Hofstede, A.H.M., Benatallah, B., Paik, H.-Y. (eds.) BPM Workshops 2007. LNCS, vol. 4928, pp. 30–41. Springer, Heidelberg (2008)
16. Casati, F., Shan, M.-C.: Semantic analysis of business process executions. In: Jensen, C.S., Jeffery, K., Pokorný, J., Šaltenis, S., Bertino, E., Böhm, K., Jarke, M. (eds.) EDBT 2002. LNCS, vol. 2287, pp. 287–296. Springer, Heidelberg (2002)
17. Grigori, D., Casati, F., Castellanos, M., Dayal, U., Sayal, M., Shan, M.-C.: Business process intelligence. Computers in Industry 53, 321–343 (2004)
18. Kassem, G.: Application Usage Mining: Grundlagen und Verfahren. PhD thesis, School of Computer Science, University of Magdeburg (2007) ISBN: 3832259953