Automated Process Model Discovery
Limitations and Challenges

Norbert Gronau and Christian Glaschke
Chair of Business Information Systems and Electronic Government, University of Potsdam, August Bebel Straße 89, Potsdam, Germany
n.gronau@lswi.de, c.glaschke@lswi.de

Keywords: Process Mining, Process Model Discovery, Screen Capturing

Abstract: The implementation of business processes through the use of information systems (ERP, CRM, PLM and MES) has become a key success factor for companies. For further development and optimization of processes, many companies haven´t trusted processes for the analysis. Surveying as-is processes is complex and only possible by manual recording. To perform this task automatically the theory shows us different approaches (process mining, Application Usage Mining and Web Usage Mining). The target of the concepts and tools is to complement the process of continuous improvement in the company with meaningful process models, which can be reconstructed from protocols and user actions in the information systems. This article focuses on the limitations of these concepts and the challenges they present and gives an outlook on how future solutions must work to speed up the process of continuous improvement and to meet the challenges of heterogeneity in IS - architectures.

1 INTRODUCTION

For more than 20 years business process management has been the leading paradigm for organizing and restructuring corporations and public entities. Although all kinds of companies use business process management in certain areas, there are some challenges that require further analysis. To name only a few:

- Improving learning while performing a business process
- Making better use of person-bound knowledge that is generated in or used during the business process
- The establishment of PDCA cycles (plan-do-check-act) in process management, that enable the detection and subsequent correction of deviations without interrupting the business process
- Typically, business processes today are supported by enterprise systems like ERP, CRM or SCM. Normally, there are deviations between the intended process covered, the ERP and the actual process that is run in the company (cf. Gronau, 2015)
- The human interface is more important than ever in most business processes, despite automation. When the automated business process is interrupted, it is a human that has to decide how to propel the process further. The description of human interfaces is by no-way interoperable now.
- A better real world awareness of the objects of business processes (persons, information, cases, instances and customer materials) currently available, as well as approaches to integrate such information into the process, is more necessary than ever.

With respect to all these new challenges, detailed and purpose-specific modeling is the precondition for a purposeful analysis of the business process necessary for its improvement. The detection of business processes and the investigation into necessary attributes of all objects tends to be very time-consuming and is still incapable of being fully automated.

There are some approaches like process mining (cf. Van der Aalst, 2011) that can help identify process patterns or recurrent instances, but the mere act of modeling itself is one of the most challenging tasks. This process also heavily influences the quality of the results. Incorrect or missing attributes

Gronau N. and Glaschke C.
Automated Process Model Discovery - Limitations and Challenges.
DOI: 10.5220/0005885100190025
In *Proceedings of the Fifth International Symposium on Business Modeling and Software Design (BMSD 2015)*, pages 19-25
ISBN: 978-989-758-111-3

or objects mean that the purpose of analysis and the goal of the improvement cannot be reached.

Therefore, this paper describes ongoing research activities to determine an approach to automatically identify business processes and model them from the information that can be derived from information systems like ERP and CRM.

2 PROCESS MODEL DISCOVERY IN TIME FROM INTERNET OF THINGS

In this section, the topic of process model discovery in the area of a total digital integration is discussed. Current trends, such as the internet of things, industrial internet and digitalization (cf. Kagermann, 2014) are some of the main drivers for the development of new concepts and technologies in the area of business process modeling. At the center of these approaches lies the question of how it is possible to attain more efficient and transparent business processes. Hence, there is great demand for a current and trustworthy as-is process model (Houy et. al., 2011). Such a model is necessary to decide how to optimize or reengineer the process. Inquiries to determine as-is processes are very complex and labor intensive, and are typically carried out by manual forms of observation and data collection.

The main question of the present paper is: How can as-is processes in a corporation or a public entity be determined automatically for further analysis? This question has tackled in the past by using a variety of different viewpoints. For instance, there are techniques that use either certain properties of technologies (Web Usage Mining, cf. Zhong, 2013) or log files from enterprise systems to reconstruct as-is processes. This approach is called Application Usage and Process Mining. These tools and methods have to be integrated into what are today more heterogeneous application landscapes with variable technologies and application systems (cf. Huber, 2015). This trend will be even more intense in the future when more system elements from the Internet of Things are incorporated within the application landscape. Given that this development will occur in the near future, the research question of this contribution can be stated as follows: What kind of information about the environment and the enterprise are necessary in order to be able to discover processes automatically? To answer this question, current research approaches are described and their limitations analyzed. Additionally, solutions that allow for the use and analysis of the available data are outlined, and the existing challenges facing the development of a new method are illustrated. At the end, an outlook for further research is given.

3 EXISTING PROCESS DISCOVERY APPROACHES AND THEIR LIMITATIONS

A well-known approach for process discovery is the concept of process mining that was developed by Van der Aalst and his research group at Technical University of Eindhoven (The Netherlands). This approach uses log files from application systems (for instance ERP systems) to reconstruct processes. To be successful in that effort, the application system has to deliver the needed information in a specific manner (as shown in table 1).

Table 1: Example for a logfile.

PID	Activity	Worker	Timestamp
452	registration	55	2011-12-24, 11:10:21
452	investigation	56	2011-12-24, 11:15:21
452	consulting	33	2011-12-24, 12:17:10
452	dismissal	55	2011-12-24, 12:47:11
453	registration	55	2011-12-24, 11:16:35
453	investigation	56	2011-12-24, 11:27:12
453	consulting	12	2011-12-24, 11:52:37
453	dismissal	55	2011-12-24, 11:59:54
454	registration	55	2011-12-24, 11:11:21
454	investigation	55	2011-12-24, 11:15:21
454	registration	56	2011-12-24, 12:17:00

An important component of this listing of process instances is the process identification number (PID). This number is used to create a process diagram based on more than one process instance (cf. Van der Aalst, 2012). In the background, petri networks are used to be able to generate process diagrams, to describe the different conditions of the process and to create a graph for visualization and analysis (cf. Van der Aalst, 2011 & Accorsi et.al. 2012).

Another approach that uses more than one input source is Application Usage Mining. In this approach, the log files are complemented by additional data from the data-base of the information system: for instance on the users, workflows and functions of the system. This information is used to further enrich the models. In the background of this approach petri networks are also used to describe the logic of the process and to summarize the possible states of a system (cf. Kassem, 2005).

Both approaches rely on the assumption that a system that is able to generate data about business processes and to deliver log files in the necessary quality exists. This assumption is not valid for enterprises with a huge number of information systems, for instance best of breed solutions, federated ERP systems or combinations of different system classes (PLM and ERP and DMS and MES) that are used together in the business process. Furthermore, customizing the application systems is very time consuming and a very high level of expertise is required. One feature that complicates this approach is that the functions executed by the application system are used as a first hint for classification and instance creation. Hence, the relevant process tasks and functions have to be known before the customizing of log files can take place.

Another possibility to record the interactions between user and system are log files of web-based systems. This approach is called Web Usage Mining. Here the access point for the process recording is the technology used (PHP, HTML, Javascript, ...). As in the process mining approach, the server can record user requests and results into logs. The session ID is used as process ID (cf. Bhart, 2014). This principle is typically used for analyzing the user‘s behavior and not for reconstructing business processes, but an adaptation for web-based application systems seems to be at least possible.

Starting from the preconditions that for a successful execution of business processes more than one application system and different technologies are used and that some of the tasks are performed outside of the information systems, a new approach is necessary to be able to automatically discover and record business processes.

To sum up, it can be said that existing approaches for the data collection either have specific knowledge about the process or concentrate on only one application system or one single technology. As information about real objects, locations and movements are not taken into account, these methods can only achieve a limited degree of accuracy. Hence, we formulate another research question: How can the user and his surrounding likewise be taken into account for process model discovery?

4 IDENTIFIED POSSIBILITIES FOR DATA CAPTURING

This section shows possibilities to collect data from performed processes without being dependent on specific application systems or technologies. Another aim of this section is to describe which information user and environment can deliver and how this information can be captured.

4.1 Screen Capturing and Optical Character Recognition

A main point of criticism about the approaches presented in section 3 is that a lot of work is necessary to configure the log data, and that meta information for the enrichment of log files in a structured manner has to be available. To obtain this information without any knowledge about the business process, an approach is needed which works independently from application systems. Application systems are ideally closed systems with complex interactions with clients that use different data formats.

There is only one object that shows all the interactions between user and system, the screen. To be able to use this valuable source of information, screen captures can be analyzed by an OCR software (Shekappa et. al., 2015). From this information a matrix can then be derived (table 2). The lines are the clients where the software captures the screen content. The columns are the different points in time when the screen capturing took place. Every cell is the result of an OCR recognition f(t) at a certain moment (for instance t1). Available screen capturing software is also able to capture the cursor positions and mouse clicks of the user (cf. Huang et. al, 2011 & Johnson et. al., 2012). Therefore, it is also possible to analyze table cells in an application system selected by the user. This approach delivers data about the interactions of the user with a terminal that is sorted by date but unstructured. It is also possible to find out the cursor position,

Table 2: Results table for OCR recording.

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13
Client 1	f(t1)	f(t2)	f(t3)	f(t4)	f(t5)	f(t6)	f(t7)	f(t8)	f(t9)	f(t10)	f(t11)	f(t12)	f(t13)
Client 2	f(t1)	f(t2)	f(t3)	f(t4)	f(t5)	f(t6)	f(t7)	f(t8)	f(t9)	f(t10)	f(t11)	f(t12)	f(t13)
Client 3	f(t1)	f(t2)	f(t3)	f(t4)	f(t5)	f(t6)	f(t7)	f(t8)	f(t9)	f(t10)	f(t11)	f(t12)	f(t13)
Client 4	f(t1)	f(t2)	f(t3)	f(t4)	f(t5)	f(t6)	f(t7)	f(t8)	f(t9)	f(t10)	f(t11)	f(t12)	f(t13)
Client 5	f(t1)	f(t2)	f(t3)	f(t4)	f(t5)	f(t6)	f(t7)	f(t8)	f(t9)	f(t10)	f(t11)	f(t12)	f(t13)

and to analyze which fields were selected by the user and which functions were performed.

4.2 Operating System Information

Another point of criticism concerning the approaches presented in section 2 is the exclusive focus on application systems. Other software running on the computer also delivers data that can be used for the reconstruction of processes and the generation of meta information.

Aside from the user entries, the operating system can also deliver other valuable information. This might include log-in credentials, the program used and the data entered. Additionally, the operating system is responsible for file operations and network access (cf. Tanenbaum, 2003). Using this information it is possible to determine which file was opened or processed.

4.3 Process ID in Application Systems

The process ID makes a substantial contribution to the discovery of a process. It makes it possible to distinguish between different process instances. From that structure, in turn, a process model can be reconstructed. In application systems, a huge set of distinct numbers for different kinds of data and levels of detail exist. Execution data exist for the accomplishment of business processes. This data, for instance, reveals the change of storage data when a delivery document changes the stored amount of an item by a booking process. This delivery document has a unique number and also points to its logical predecessors. The offer number can be found in the order; the order number is mentioned on a factory order or on a delivery document. The document flow that can be created from these numbers can be reconstructed or read from the leading application systems supporting this process. Additionally, these numbers are also available when corresponding with a client, for instance ticket numbers or invoice numbers. Master data identifies a business object (product resource, storage facility) uniquely by a number. In some cases, not only are the products identified, but also the object instances. Therefore, a serial number or batch number is used. These numbers are typically printed onto the product and are used for traceability of every single item or batch job in logistics and manufacturing.

4.4 Tracking of Business Objects

Another approach to collect information about the environment and the user utilizes technologies to spatially locate users and objects. Different approaches, such as RFID or GPS, exist for this purpose. A movement profile can be derived from this data. These movements can be used to follow processes when an object is moved and to define interactions between users and business objects (cf. Sultanow, 2015 & Gronau, 2014).

4.5 Summary

Four approaches can be used to collect data about business processes in corporations. The next task is to use that information to reconstruct the underlying business processes.

5 CHALLENGES DURING THE DEVELOPMENT OF A METHOD

Combining the information from the different approaches in section 4 leads to different challenges for further research. In this section, the problems are elaborated with the help of an exemplary scenario. This scenario is described at first to depict the challenges. A special emphasis is put on the aspects of the existing approaches Process Mining, Application Usage Mining and Web Usage Mining that constitute weaknesses.

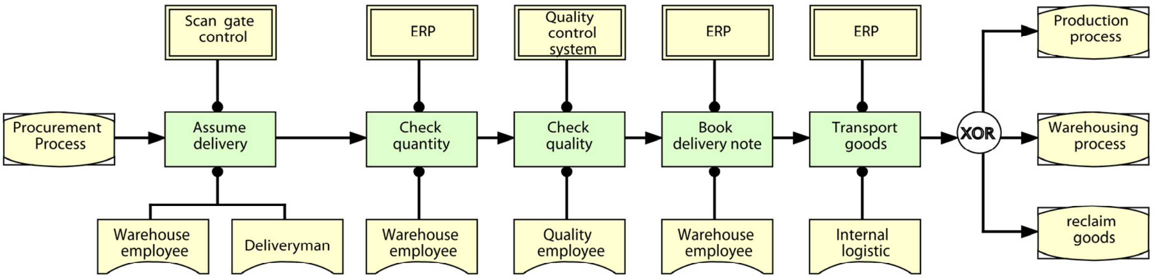

Figure 1: Product entry process.

5.1 The Example of the Data Entry Process

To be able to describe the challenges a scenario was developed, in which a product entry process is performed. In figure 1, the process is illustrated in the KMDL process view (Gronau, 2005).

The process goes as follows: after a successful order from the procurement process (process interface), the goods are delivered. The delivery is accepted and driven through a scan gate. The information system „scan gate control“ shows the warehouse employee (role) the listing of the delivered parts on a client computer and shows the serial, order and supplier numbers from the RFID tags of the packages. After checking that information, the warehouse employee approves the delivery document for the delivery man and in the scan gate control Software the parts list. The data from the scan gate control system is entered into the ERP system as a quality control system. In the ERP system, a new bill of delivery is generated that contains the articles and their amounts. The warehouse employee unpacks the goods and gets the bill of delivery in the ERP system in order to check the quantities. After confirming the amount, the ERP bill of delivery is handed over for inspection to the Quality Control System, which creates a test order, which includes the different properties of the articles. The Quality employee checks the properties per serial number and deposited a test result. Some parts are identified as unsuitable. This has to be entered into the quality control system. Most articles pass the amount check successfully. Acknowledging that the ERP system has been checked automatically produces the results of the quality assurance task and generates a storage location for every serial number of the delivered goods in the bill of delivery. Good parts are now transported to the manufacturing storage, bad parts into a reclaim storage yard. Articles marked as consumables are stored in the warehousing process. In process step book delivery note the articles in different charge carrier / transport units are separated. Then, the delivery will be booked. The ERP system generates from a transfer order, which is processed by the logistics department. Hence, the goods are stored in different locations.

5.2 Challenges of Automated Process Model Discovery

The first challenge for the approaches from section 3 is to recognize that the procurement process is the trigger for the warehouse entry process. In the process “assume delivery,” the order number is the only connection between these two processes. This connection has to be recognized and assigned to the order process. This is the main result to expect from the method to be developed. Therefore, the recognition mechanism has to find out that the same order number is now used during the delivery, and has to interpret this as a unique number.

Using the statements above we can formulate some requirements for the necessary recognition mechanism.

The method has to reliably find out the unique identifiers by screen capturing and OCR.

A second challenge can be derived from the “assume delivery” process. Here it is necessary to find out that the serial numbers indicate the different flows of goods. To achieve this, the products that are equipped with RFID tags at different locations (storage, quality control, ...) show the serial numbers and the current location belonging to that serial number. Of course, the information about the locations of the RFID reader stations must be known. Second challenge: The method must combine information from the screen capturing with the determined location of the goods.

Another challenge is the assignment of roles and information systems to the steps of the process. The warehouse employee logs himself into the system with his mobile device and connects himself to the scan gate control system. The operating system then captures the user group or the log-in name. With other information systems the process is performed in a similar manner. The operating system registers the usage of the scan gate control software and allocates this software to the process. The only role in the exemplary process that cannot be captured in this manner is the delivery man. Third challenge: The method must be able to recognize the external and internal roles involved in the process.

The next challenge is it to differentiate between the different process steps. In the example, the differentiation between the process steps “assume delivery” and “check quantity” can be performed by different information systems. When the process transfers from “check quantity” to “check quality”, different roles and different systems allow one to find out that different process steps are performed. A differentiation on that level is at any rate impossible. An example of this can be seen in the work of a sourcing employee who works with an ERP system and does everything in the sourcing process, from ordering to invoice checks, on his or her own. Fourth challenge: The method must discover the different process steps and be able to see the limit of one process step and the beginning of another one.

The fifth challenge is to find out the description of the process. To that end, a lot of information is collected from OCR or screen capturing, but their interpretation is difficult. An example is the “check quantity” process, and the question: how can we derive that term? One approach would be to assign the function to the location; another approach would be to use the window title of the ERP system (“delivery note”). Sometimes this task can be done by manual configuration, or by screen capturing. The fifth challenge is, therefore, that the method must be able to determine the name of a process step.

The sixth challenge is to recognize different target locations (storage locations in the logistic process) from logistics and from the transport of goods. Therefore, these different locations have to be distinguished in the process by using different process interfaces. To meet this challenge, the master data of the storage in the warehouse management system could be used to help understand the structure of the storage groups and their functionality in the process. Another possibility is to assign this information to the different locations. When this information is available for differentiation, the process interfaces into the storage area can be reconstructed. The sixth challenge is: The method must have knowledge that specifies the environment.

6 CONCLUSION AND OUTLOOK

The contribution has shown that an automated discovery of process models is possible when some new approaches are applied. The investigation of current approaches showed that systems and technologies deliver valuable information about the process flow, but a configuration for a case of specific use is necessary. The main gap in the research is the lack of consideration of human tasks and environmental data.

For the research task to develop a new integrated approach, a couple of challenges must be dealt with. One of the most important requirements of a new method is to see the corporation and its data sources in an integrated manner. Another important topic is the collection of data according to location, time and their connection to the process model. No satisfactory answer could be given to the research question concerning which information about the user and the environment has to be collected in order to be able to sufficiently discover process models. On one hand, information about location must be available (for instance which task is performed where), while on the other hand, the master data that holds that information has to be investigated. In any case, the demand for and benefit of that kind of input can be shown. Finally, there remains the question of how the recognition mechanism uses semantic techniques. Here it might be possible that the user has to assist the recognition mechanism to describe the process models.

An open issue after creating process models is to interpret these semi-formal models. To reach an understanding about a process solely by using a model is very difficult. The authors thinks that human beings, too, will have to participate in that process in the future.

REFERENCES

Accorsi R., Stocker T. (2012). *On the Exploitation of Process Mining for Security Audits: The Conformance Checking Case*. ACM Symposium on Applied Computing. doi:10.1145/2245276.2232051

Bhart, P. (2014). *Prediction Model Using Web Usage Mining Techniques.* IJCATR, Volume 3, Issue 12, 827-830. doi: 10.7753/IJCATR0312.1015

Gronau, N. (2015). *Trends and Future Research in Enterprise Systems.* Lecture Notes in Business Information Processing, Volume 198, 271-280.

Gronau, N., Müller, C., & Korf, R. (2005). *KMDL – Capturing, Analysing and Improving Knowledge-*

Intensive Business Processes. Journal of Universal Computer Science, 11(4), 452-472.

Gronau, N., Sultanow, E. (2014). *Echtzeitmeldung und Analysen über Wissensereignisse. IM+io Fachzeitschrift für Innovation, Organisation und Management*. 01/2014. 80-87

Houy C., Fettke P., Loos P., Van der Aalst WMP., Krogstie J. (2011). *Business process management in the large*. Business Information System Engineering 3, 385-388. doi:10.1007/s12599-011-0181-5.

Huang, J., White, R. W., Dumais, S. (2011). *No Clicks, No Problem: Using Cursor Movements to Understand and Improve Search.* Proceeding CHI '11 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1225-1234. doi: 10.1145/ 1978942.1979125

Huber, S., (2015). *Informationsintegration in dynamischen Unternehmensnetzwerken – Architekturen, Methoden und Anwendungen*. Springer. doi: 10.1007/978 -3-658-07748-8

Johnson, A., Mulder, B., Sijbinga, A., Hulsebos, L. (2012). *Action as a Window to Perception: Measuring Attention with Mouse Movements.* Applied Cognitive Psychology, Appl. Cognit. Psychol. 26, 802-809. doi:10.1002/acp.2862

Kagermann, H., (2014). *Chancen von Industrie 4.0 nutzen. Industrie 4.0 in Produktion, Automatisierung und Logistik*. Springer. 603-613. doi: 10.1007/978-3- 658-04682-8

Kassem, G., Rautenstrauch, C. (2005). *Application Usage Mining to Improve Enterprise Workflows: ERP Systems SAP R/3 as Example*. IDEA Group Publishing, In:Proceedings of the 2005 Information Resources Management Association International Conference.

Krogstie, J. (2015). Capturing Enterprise Data Integration Challenges Using a Semiotic Data Quality Framework. BISE, 57(1). 27-36

Schemm, J. W. (2009). *Zwischenbetriebliches Stammdatenmanagement*. Springer Verlag.

Shekappa B., Mallikarjun, A., Shivarama, J. (2015). *Best Practices in Digitization: Planning and Workflow Processes.* In International Conference on the theme Emerging Technologies and Future of Libraries: Issues and Challenges. 332-340

Sultanow, E., Cox, S., Brockmann, C., Gronau, N. (2015). *Real World Awareness via the Knowledge Modeling and Description Language*. In M. Khosrow-Pour (Ed.), Encyclopedia of Information Science and Technology, Third Edition, 5224-5234. doi:10.4018/978-1-4666-5888-2.ch516

Tanenbaum, A. S. (2003). *Moderne Betriebssysteme.* Pearson Studium, Auflage 2.

Tiwari, A. Turner, C.J., Majeed, B. (2008). *A review of business process mining: state of the art and future trends.* Business Process Management Journal. Vol. 14 Iss: 1. 5-22.

Van der Aalst, WMP. (2011). *Process Mining – Discovery, Conformance and Enhancement of Business Processes*. Springer.

Van der Aalst, WMP., Accorsi R., Ullrich M. (2012). *Process Mining*. Retrieved April 29, 2015, from http://www.gi.de/nc/service/informatiklexikon/detailansicht/article/process-mining.htm

Zhong, N., Liu, J.,Yao Y. (2013). *Web Intelligence. Web Log Mining*. Springer. 173-198.