A Review of Process Discovery Methods and Conformance Checking Methods

Yuheng Zhang[1 a], Yi Zhang*[2,3]

[a] 805267696@qq.com; * zhy_gsdx@163.com

[1] Chongqing University of Science and Technology, Chongqing, China,
[2] Institute of Chengdu-Chongqing Economic Zone Development, Chongqing, China
[3] Regional Economic Research Institute, Chongqing University of Technology and Business, Chongqing, China

Abstract. With the ascent of data science, process mining has garnered increased attention. The objective of process mining is to extract valuable insights from event logs, facilitating the discovery, monitoring, and enhancement of real business processes. Process mining is primarily categorized into three research areas: process discovery, conformance checking, and process enhancement. The aim of process discovery is the automated extraction of process models from event logs. Conformance checking is primarily employed to assess the quality of the extracted models, thus evaluating the effectiveness of process discovery methods. Process enhancement involves expanding the model based on the outcomes of conformance checking. Conformance checking is primarily categorized into four quality dimensions: fitness, precision, generalization, and understandability. This paper predominantly examines process discovery and conformance checking methodologies.

Keywords: Data mining; Process mining; Conformance Check; Petri Net

1 Introduction

Compared with traditional data-centric methods (such as data mining) and process-centric methods (such as business process management analysis), process mining not only analyzes the data, but also analyzes the end-to-end process, so it can better analyze the operation of the enterprise. Process mining techniques can examine when and how processes deviate from the designed process model, as well as identify activity bottlenecks in business processes and what causes delays in product delivery and enterprise services. Process mining has proven successful in many fields, where it helps with challenging tasks such as fraud detection, robotic process automation, or learning analysis. In addition, process mining is also favored in the analysis of some processes that need to ensure safety, such as in the field of healthcare, which can provide treatment procedures and patient treatment plans in critical situations. Through data visualization components and process mining software, users can dig deep into the data to discover deviations between actual and expected processes, as well as the root causes of inefficiencies in business operations

Nowadays, business processes in enterprises are becoming increasingly complex and more difficult to draw manually. For enterprises, how to obtain a high-quality business process model is becoming increasingly important. Extracting business data information recorded in

BDEDM 2024, January 12-14, Ningbo, People's Republic of China

DOI 10.4108/eai.12-1-2024.2347216

traditional information systems and processing it, converting it into event logs that can be input into process mining methods. The purpose of Process Mining is to automatically extract useful information from event logs generated by information systems, thereby discovering, monitoring, and improving actual business processes. Process mining mainly has three application areas: process discovery, conformance checking, and process enhancement [1].

Process mining not only establishes a link between the actual process and data but also connects various models. The initiation of process mining occurs with the event log generated within the information system, as illustrated in Figure 1. This figure also outlines the three primary application fields of process mining.

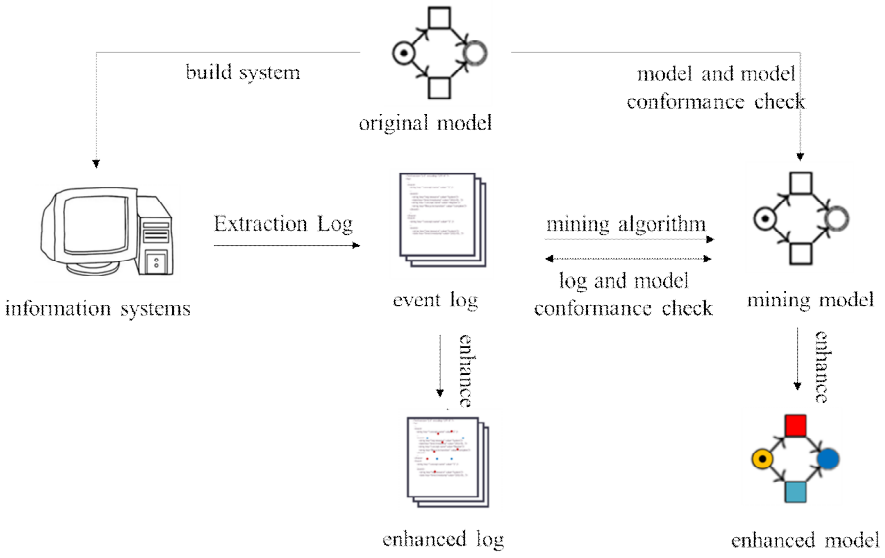

Figure 1. Application areas of process mining.

The purpose of process discovery is to automatically extract process models from event logs. The process model obtained by mining event logs through process discovery methods is called a mining model. The main aim of conformance checking is to evaluate the results of process discovery, that is, the quality of the mining model, thereby evaluating the process discovery methods. Conformance checking is mainly divided into two parts: conformance checking of logs and models, and conformance checking of models and models. Conformance checking of logs and models compares and analyzes event logs and mining models, and conformance checking of models and models compares and analyzes original models and mining models. Conformance checking can reflect whether the process model conforms to the expected design, thereby reflecting the strengths and weaknesses of the process discovery. Process enhancement is to improve or expand the process model or event log, so that it can better conform to the actual process information. When the process model cannot accurately reflect reality, the value of process enhancement becomes apparent, such as enhancing the model by adding new perspectives, and improving the adaptability of the model.

A process is a series of steps or stages through which certain tasks can be accomplished. For example, manufacturing processes include casting, forging, stamping, welding, machining, and assembly. A log is a record of activities generated during the operation of an information system, and each recorded activity is related to the execution of the process. Nowadays, information systems can be found everywhere, and more and more log information can be recorded. Therefore, the process information needed to achieve a certain purpose can be extracted from the log, thus forming the event log required for process mining. With the vigorous development of process mining at home and abroad, process mining technology is becoming more and more mature. Process mining can help enterprise managers monitor the actual operation of enterprises, thus assisting enterprises to intelligently obtain the actual process, compare the actual process with the expected process, analyze the work efficiency of enterprises, detect the bottleneck of enterprise operation process, and provide suggestions for process improvement. With the increasing size of event logs, the process model structure obtained by process mining becomes more and more complex, such as repetitive tasks, implicit repositories, invisible tasks, cyclic structures, non-free choice structures, etc.. These complex structures increase the difficulty of process mining. Most of the current conformance checking methods are limited and cannot reasonably and effectively evaluate process models that contain repetitive tasks. Therefore, it is an important research on how to carry out consistency check on the process model with repetitive tasks.

2 Current research status of process discovery methods

The relevant theoretical research of process mining has been developed for decades, among which W. Van Der Aalst and his team members from the University of Eindhoven in the Netherlands have made more contributions in the field of process mining, and his team has also developed the Prom[2] open source tool. In literature [3], W. Van Der Aalst et al proposed α algorithm, which is one of the most widely used process discovery algorithms. The algorithm can mine the process model based on the order and dependencies between the activities in the event log. However, it does not effectively deal with noise in logs and complex structures in models. Therefore, many researchers have made further improvements to the α algorithm. A. K. Medeiros et al. [4] proposed an α^+ algorithm for mining short cycle structures. In reference [5], α^{++} algorithm is proposed, which can mine non-freely selected structures and implicit repositories between activities. An $\alpha^\#$ algorithm is proposed in reference [6], which can effectively deal with invisible tasks in the model. There are also some improvements based on α algorithm to make it possible to mine repeated tasks in the model, such as α^* algorithm in reference [7] and τ algorithm in reference [8].

To address the noise issue in the α algorithm and its extension, some people [9] introduced the Heuristics Miner (HM), an extension of the α algorithm. By considering the frequency information of the relationship between activities in the event log, HM can mine a broader set of process model constructs, effectively eliminating noise in the log. Although the process model produced by this algorithm is more reasonable than the previous one, it struggles with repetitive tasks and invisible tasks. A. Burattin et al. [10] proposed an improved Heuristics Miner algorithm, which is capable of processing stream event data. The method uses a hierarchical experimental design, a structure that grants researchers complete control over the behavior in the corresponding event data.. Some people [11] developed an algorithm based on

log clustering, which is optimized on Heuristics Miner algorithm, which identifies different modification of the process model by collecting some log tracks which is similar. It performs hierarchical clustering on the log, where each track is treated as a point in the identified feature space, and the resulting model is a disjunction pattern that explicitly deals with variations of the process. Fodina algorithm [12] is also an improvement based on Heuristics Miner algorithm. The process model obtained by this method is of better quality, and the method can find repetitive tasks. There is another extension of the Heuristics Miner algorithm in literature [13], which incorporates the timestamp information of an activity based on the HM algorithm, and represents the activity as a period of time instead of a single event.

Other process discovery technologies use heuristic networks to represent the mining process model. The most famous one is Genetic Algorithms (GA) [14], which is an adaptive search method by imitating the evolution process of nature. It regards the process model as the initial population of individuals and uses crossover or mutation and other means. Combine the activities into a new model. In literature [15], Evolution Tree Miner (ETM) is proposed. ETM is an improvement of genetic algorithm, which can optimize the evaluation results of mining model in four quality dimensions of consistency check.

In the process discovery method that can mine repetitive tasks, except Fodina algorithm, Artificial Negative Events algorithm [16] (Artificial Negative Events, AGNEs), Splitting Labels After Discovery (SLAND) algorithm [17] (Splitting Labels After Discovery (SLAND)) can also excavate repetitive tasks. AGNEs method represents process mining as a problem of multi-relation classification learning of event logs, and adds artificially generated negative information to improve model quality. The SLAND method utilizes local information from the log to improve the previously mined model, conducting local searches for tasks in the log that are more likely to exhibit repetition.

3 Current research status of conformance check

The purpose of conformance checking is to compare event logs and process models to detect process deviations and obtain diagnostic information [18]. The reason for these biases is related to the process execution not following the process model (for example, the execution of some activities may be missing, or the activities do not occur in the correct order), so the behavior observed in the log needs to be correlated and compared with the behavior observed in the model.

It is also difficult to assess the quality of the process models associated with logging. These are often referred to as the four quality dimensions of process mining. The four dimensions of consistency check are: adaptability, accuracy, generalization and comprehensibility[19]. Over the years fermentation, For each quality dimension in the conformance checking method, there are many methods [20], as shown in Figure 2.

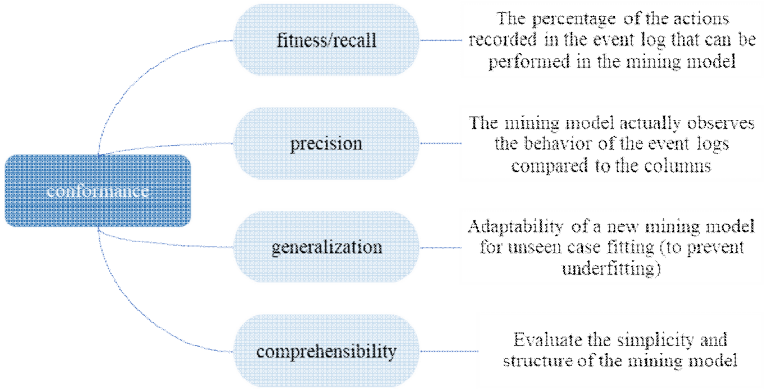

Figure 2. Four dimensions of conformance check.

3.1 Fitness

Fitness (also known as fitness/recall) represents the proportion of the number of tracks in the event log that can be successfully executed in the mining model to the total number of tracks. For example, some people[3] proposed the footprint matrix method, which mainly arrests the causal relationship between different activities in the log. A. J. Weijters et al. [9] proposed the Continuous Paring Measure (CPM) evaluation method. CPM uses replay technology by transforming Petri nets into a causality matrix abstracted from the representation of the process model. This matrix defines the input and output expressions for each activity, describing possible input and output behaviors. When replaying event logs on the causality matrix, it is necessary to check whether the corresponding input and output expressions are enabled, thus allowing the activity to be executed. Although this method can assess models containing invisible tasks, it cannot perform consistency checks on models involving repetitive tasks. A. Rozinat et al. [21] proposed an adaptive evaluation method f based on Token replay technology (the method is called token-based replay method). In this method, the event log is replayed in the process model. Less the number of token lost and remaining in the process of replaying, the better the model adaptability. S. Goedertier et al. [22] used true positive and false negative counters to calculate the degree of fit. True positives represent the number of events that can be correctly resolved in the mining model, i.e. the transition triggered is enabled. False negative indicates the number of events that need to be forcibly triggered instead of triggering the corresponding transition required for the execution of the simulated event flow.

3.2 Precision

Precision represents the proportion of the behavior tracks that can actually be observed from the mining model to those recorded in the event log. This quality dimension prevents overfitting of the process model. Model overfitting refers to the general nature of the process model, which can perform behavioral trajectories that do not appear in the event log. Some people [22] developed a method that uses negative events for accuracy detection. This method mainly relies on the idea of confusion matrix in the field of data mining. In this confusion matrix, the induced negative event is treated as the real situation, and the process model is treated as a prediction machine to predict whether a certain event will occur. A negative sample is represented as a place in the trajectory where a particular event cannot occur. A.

Rozinat et al. [21] proposed the behavior fitness evaluation method α_B and the advanced behavior fitness evaluation method α'_B. α'_B is an improvement on the basis of α_B. They converted the following relation in the model into A causal dependence matrix and used footprint comparison technology to evaluate the accuracy of the model, but this method is time-consuming and overspent.

3.3 Generalization

Generalization refers to the adaptability of the new unseen case fitting mining model to prevent underfitting of the model. There are few evaluation methods for this quality dimension, and one method that can evaluate the generalization is the alignment generalization evaluation method [23]. In addition, the anti-homogeneous generalization method [24] and the g_B^W method can also evaluate the generalization of the model. The anti-homogeneous evaluation method is a weighted combination of anti-homogeneous generalization based on trajectories and logs.

3.4 Comprehensibility

Comprehensibility, which represents the complexity of the process model, can be divided into two dimensions: conciseness and structure. The purpose of conciseness is to evaluate the number of elements in the process model. The fewer elements in the model, the better the conciseness of the model, provided that the process can be described clearly. However, this quality dimension is deliberately downplayed and there are few evaluation methods. Structure is mainly to evaluate the type and number of process structures in the process model. There are too many complex structures in the model, and it is not easy for users to understand the meaning of the model, so a well-structured model should not have too many unnecessary complex results. At present, the structural fitness evaluation method α_s proposed by A. Rozinat et al. [21] and the advanced structural fitness evaluation method α'_s are used more frequently, both of which calculate the structural adaptability by calculating the proportion of complex structures in the model. They believe that the process model should not expand due to complex structures. A. Dikici et al. [25] proposed the Control Flow Complexity (CFC) method. The structuremetric (SM) evaluation method proposed in literature also had a high usage rate.

4 Conclusion

This paper mainly reviews the methods of process discovery and consistency check. As event logs become larger and larger, the structure of process model obtained from process mining becomes more and more complex, and these complex structures increase the difficulty of process mining. Most of the current process discovery and conformance checking methods are limited and cannot reasonably and effectively discover or evaluate models containing complex structures. Therefore, it is an important research on how to carry out consistency check reasonably on the process model with complex structure. At present, conformance checking methods have been studied in various dimensions, but one of the core challenges is how to balance multiple quality dimensions at the same time, so as to obtain a mining model with good quality and easy for users to understand. This paper only reviews the theory and technology of process mining, but process mining is not only a theoretical study, but also can

be applied to business process management, business management, business consulting and other practical scenarios.

Acknowledgments. Research project on the promotion path of urban-rural integrated development in the upper reaches of the Yangtze River, the project number KFJJ2022004.

References

[1] Van Der Aalst W. Process mining data science in action. Second Edition[M]. Berlin, Germany: Springer Press, 2016: 30-33, 174-177, 185-187.

[2] Van Dongen B. F., et al. The ProM framework: a new era in process mining tool support[C]// International Conference on Application and Theory of Petri Nets. Miami, FL, USA: Springer Press, 2005: 444-454.

[3] Van Der Aalst W., et al. Workflow mining: discovering process models from event logs[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(9): 1128-1142.

[4] De Medeiros A. K., Van Dongen B. F., et al. Process mining for ubiquitous mobile systems: an overview and a concrete algorithm[C]// International Workshop on Ubiquitous Mobile Information and Collaboration Systems. Riga, Latvia: Springer Press, 2004: 151-165.

[5] Wen Lijie, et al. Mining process models with non-free-choice constructs[J]. Data Mining and Knowledge Discovery, 2007, 15(2): 145-180.

[6] Wen Lijie, Wang Jianming, et al. Mining process models with prime invisible tasks[J]. Data and Kowledge Engineering, 2010, 69: 999-1021.

[7] Li Jiafei, Liu Dayou, Yang Bo. An extended α algorithm for finding repetitive tasks in process mining [J]. Chinese Journal of Computers, 2007, 30(8): 1436-1445.

[8] Gu Chunqin, standing friends, Tao Gan, etc. Process mining algorithm for Solving Complex Tasks [J]. Computer Integrated Manufacturing Systems, 2009, 15(11): 2193-2198

[9] Weijters A. J., et al. Process mining with the heuristics miner algorithm[R]. Eindhoven, The Netherlands: BPM Center Report, 2006: 1-34.

[10] Burattin A., Sperduti A., Van der Aaslt W. Control-flow discovery from event streams[C]// 2014 IEEE Congress on Evolutionary Computation (CEC), BeiJing: IEEE Press, 2014:2420-2427.

[11] Greco G., Guzzo A., Pontieri L., et al. Discovering expressive process models by clustering log traces[J]. IEEE Transactions on knowledge and data engineering, 2006, 18(8): 1010-1027.

[12] Vanden Broucke S. K., De Weerdt J. Fodina: a robust and flexible heuristic process discovery technique[J]. Decision Support systems, 2017, 100: 109-118.

[13] Burattin A. Heuristics miner for time interval[J]. Process mining techniques in business environments, 2015, 11:85-95.

[14] De Medeiros A. K. Genetic process mining[D]. Eindhoven, The Netherlands: Eindhoven University of Technology, 2006.

[15] Buijs J. C. A. M., et al. van Der Aalst W. M. P. On the role of fitness, precision, generalization and simplicity in process discovery[C]// OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", Berlin, Germany: Springer Press, 2012: 305-322.

[16] Goedertier S., et al. Robust Process Discovery with Artificial Negative Events[J]. Journal of Machine Learning Research, 2009, 10: 1305-1349.

[17] Vázquez-Barreiros B., et al. Enhancing discovered processes with duplicate tasks[J]. Information Sciences, 2016, 373: 369-387.

[18] Asare E., et al. Conformance Checking: Workflow of Hospitals and Workflow of Open-Source EMRs[J]. IEEE Access, 2020, (99): 1-1.
[19] Berti A., et al. A Novel Token-Based Replay Technique to Speed Up Conformance Checking and Process Enhancement[J]. Transactions on Petri Nets and Other Models of Concurrency XV, 2021, 12530: 1-26.
[20] Tax N ,et al.The imprecisions of precision measures in process mining[J].Information Processing Letters, DOI:10.1016/j.ipl.2018.01.013.
[21] Rozinat A., et al. Conformance checking of processes based on monitoring real behavior[J]. Information Systems, 2008, 33(1): 64-95.
[22] Goedertier S., Vanthienen J., et al. Robust Process Discovery with Artificial Negative Events[J]. Journal of Machine Learning Research, 2009, 10: 1306-1349.
[23] Van der Aaslt W., et al.. Replaying History on Process Models for Conformance Checking and Performance Analysis[J]. WIREs Data Mining and Knowledge Discover, 2012, 2(2): 182-192.
[24] Van Dongen B. F., et al. A unified approach for measuring precision and generalization based on anti-alignment[C]// International Conference on Business Process Management. Rio de Janeiro, Brazil: Springer Press, 2016: 39-56.
[25] Dikici A., Turetken O., Demiros O. Factors influencing the understandability of process models: A systematic literature review[J]. Informantion and Software Technology, 2018, 93: 112-129.