Implementation of Genetic Process Mining to Support Information System Audit

[1]Yora Radityohutomo, [2]Gede Agung Ary Wisudiawan, [3]Andry Alamsyah, [4]Anisa Herdiani
[1,2,4] School of Computing, [3]School of Economics & Business, Telkom University
Bandung, Indonesia
[1]yoratyo@gmail.com, [2]degunk@telkomuniversity.ac.id, [3]andrya@telkomuniversity.ac.id,
[4]anisaherdiani@telkomuniversity.ac.id

Abstract

One of the frameworks that can be used to audit information systems is COBIT 5 which offers *process assessment model* (PAM). The process assessment model usually done by collecting and validating random factual data samples, so that the results of this assessment cannot be representative of the overall ongoing process. This research uses *process mining* by using event log to replace data collection and data validation stage in *process assessment model*. *Process mining* aims to describe the ongoing process model of the event log data automatically so that it can be compared with the standard flow process in real time. *Process mining* is applied using a *genetic algorithm* that can recognize less frequent behavior in event log as noise data. This assessment process delivers the rating point level as a result for comparison of the standard process flow with the process model of the process mining and business flow analysis of the event log data. The results of this study show *genetic process mining* able to support corporate information system audit activities.

Keywords: Process Assessment Model; Process Mining; Genetic Algorithm

INTRODUCTION

Today almost all large companies perform the monitoring function of the company's business processes with an information system audit to ensure the performance and quality of the business process implementation of the company runs in accordance with the planning and business objectives [1]. Information system audit can be done by assessment process in accordance with IT governance standard framework. One of the frameworks that can be used is COBIT 5 which provides a *process assessment model* (PAM) to test the capabilities of IT processes [2].

Process assessment model consists of planning phase, data collection, data validation, process attribute rating and reporting. In the data collection stage, the assessor collects data about the process, which includes the input, output and objectives of a process to support assessment [2]. The data collection is done by taking a random sample of factual data, so that a lot of unrepresented data and the data collected are subjective [1]. Then in the data validation stage the assessor ensures that the data is accurate enough and covers the scope of the assessment by validating the collected information [2]. Validation process takes a long time because it needs to be done repeatedly when there are data changes in the running process, so the results cannot be obtained in real time.

We can use *process mining* method which store their event log to handle this problem. *Process mining* can describe the running process model of all event log data automatically so that the results can be analyzed in real time [3]. This study uses a process assessment model that implements *process mining* to replace the data collection and data validation stage [2]. *Process mining* is applied using a *genetic algorithm* [6], which is a process model search technique following the principle of evolution that the quality of the process model is judged by comparing it to all traces in event log [4]. So, the resulting solution is global and can handle the problem of event log data containing noise [4]. The process assessment model in this research is applied to the domain DSS01 (Manage Operations) on DSS01.01 practices (perform operational procedures), because only in this domain allows for the implementation of *process mining* [2].

This research applies to case study of information system audit in distribution company. The result of process assessment model in this research is level rating point as result of comparison of standard flow process with process model from *process mining* and business flow analysis from event log data.

THEORY

2.1 Process Mining

Process mining is one of the techniques developed based on data mining, which is the difference is the process of mining focus on the activities that occur. Processed data is event log extracted from the activity that occurs. Event log are data that contain information that can describe the behavior of processes that occur [3]. Event log can be obtained from information systems and data from the units involved in the process. Not all data can be used as event log, not all information is also required in event log. The data in the event log must contain information that contains a set of events and cases sorted by time so that the data can describe a running process [3].

The main purpose of *process mining* is to process the event log into a model to provide recommendations to the process model. There are three stages of *process mining namely Process Discovery, Conformance checking,* and *Enhancement* [4] that can be seen in figure 1.

Process discovery is a process model stage of the data that can be from event log, so it can describe the process model in accordance with the real process in the field. The process of *conformance checking* is the phase comparison between the process model of *discovery* with all trace event log to see the suitability of the model construction. The latter is the stage of *enhancement*, which is the stage for developing and recommending a process model of an existing process model and has been adapted to process modeling of event log data [3].

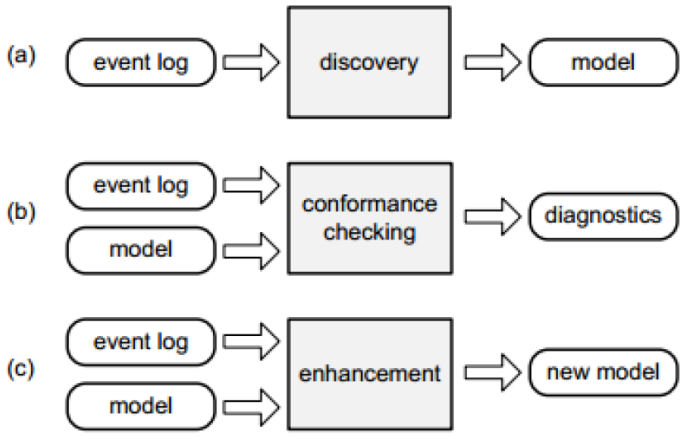

Figure 1. Process Mining Stages [3]

2.2 Genetic Process Mining

In *genetic process mining*, the individual is a process model, a *fitness* value (1) that measures how well an individual (or process model) reflects behavior in event log, and genetic operators recombine individuals so that new process model candidates can be created.

$$fitness = 0.40 * \frac{allParsedActivities}{numberOfActivitiesAtLog} + 0.60 * \frac{AllProperlyCompletedLogTraces}{numberOfTracesAtLog} \quad (1)$$

Therefore, the challenge is to define an internal representation that supports all common constructs in the process model including sequence, parallelism, choice, repetition. Genetic algorithms can handle *noise* data such as duplicated tasks, missing heads, missing bodies and missing tail because *fitness* measures are measured by replaying all event log data to individual process models [3]. To assess the quality of process models created (or individually) in each population and genetic operator so that all space searches can be defined with internal representations can be explored [4].

Method

An overview of the system can be seen in Figure 2 which refers to the COBIT 5 model assessment process described in the previous section.

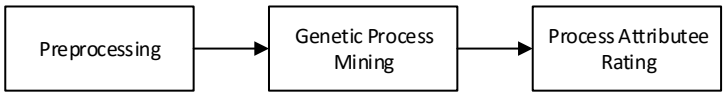

Figure 2. Out Proposed General Flow System

3.1 Preprocessing

The event log data will be processed beforehand through preprocessing to simplify and equate the format to fit the system design. The event log data used in this study belongs to the distribution company [7]. Preprocessing is done by removing some unused columns. The data needed by the system is the case id, the name of the activity, and the time and executor of the activity.

3.2 Genetic Process Mining

The *genetic process mining* stage is described in detail in the flow chart of Figure 3.

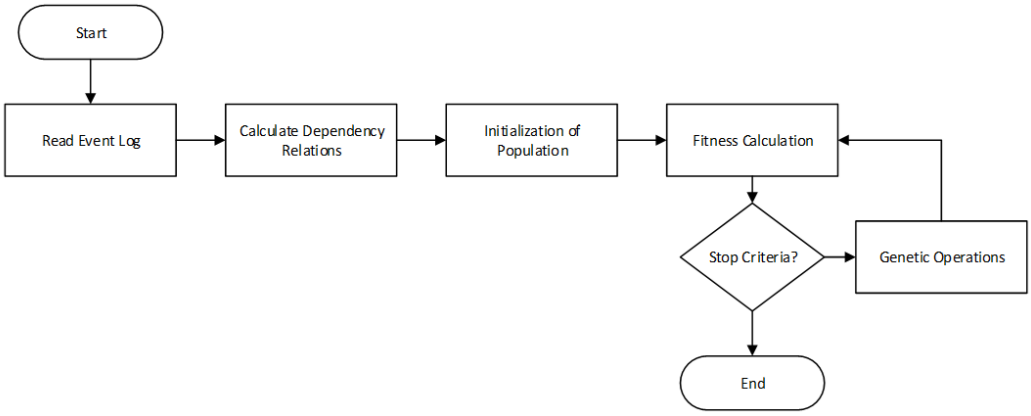

Figure 3. Genetic Process Mining

The results of *genetic process mining* are the discovery model process and bottleneck analysis from event log data. Discovery model process is constructed with genetic algorithm that has four parameters, they are *population size, maximum generation, crossover probability* and *mutation probability* [6].

3.3 Process Attribute Rating

This process attribute rating includes the process of comparing the process models of *genetic process mining* with the standard process model. This comparison is done using set theory, which is the ratio from the number of intersection and number of union between the two model processes. The similarity of two models get the capability level that can be seen in table 1.

Table 1. Capability Level of Process Attribute Rating [5]

Level	Achievement
N: Not Achieved	0 % - 15 %
P: Partially Achieved	15 % - 50 %
L: Largely Achieved	50 % - 85 %
F: Fully Achieved	85 % - 100 %

Result & Analysis

Based on the result of testing the *population size* and *the maximum generation* then the selected parameter value is at the time of the *population size* = 40 and *maximum generation* = 20, because at this parameter value the *fitness* value reaches the highest before finally entering the saturation period resulting in the increase of the *fitness* value is not too significant. This decision also considers the computation time that will be greater if the population and maximum generation is greater.

Based on the result of *crossover probability* test and *mutation probability* then the selected parameter value is when *crossover probability* = 0.9 and *mutation probability* = 0.1, because at this parameter value the *fitness* value reaches the highest. These results are in proportion to research on genetic algorithms that the odds of crossing should be high in order to exchange solutions between individuals, whereas mutation opportunities should be low in order not to damage the quality of the individual.

4.1 Result of Process Attribute Rating

After obtaining the value of the appropriate genetic parameters that is with the value of *fitness* 0.932 on the model 1 and 0.91 in model 2, then the results of both models with the best *fitness* process will be compared with the standard flow process of distribution company shown in Figure 4.

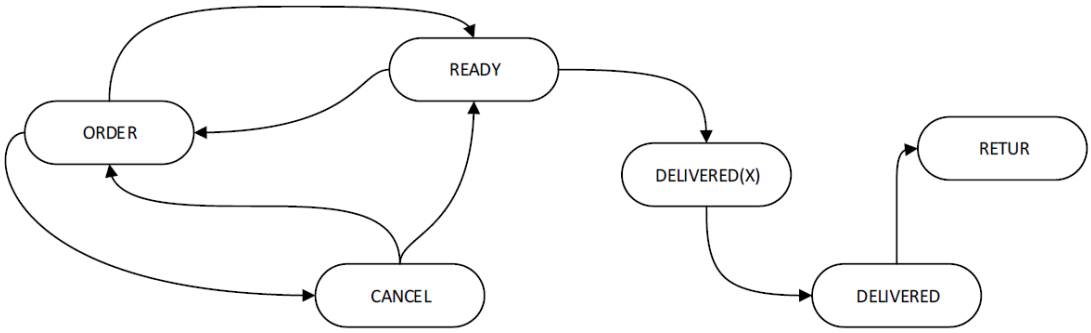

Figure 4. Standard Process Model

The shape of the process flow from model 1 (m1) can be seen in Figure 5 When compared with the standard process model then get the rating point value level of 63.64% to obtain L (Largely Achieved) capability level.

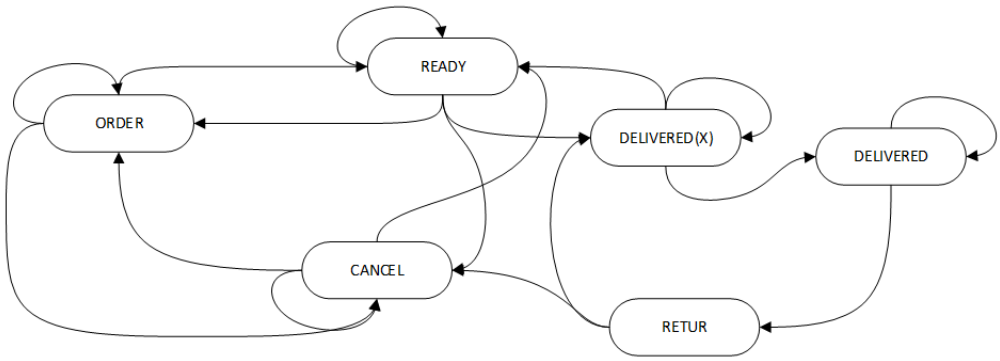

Figure 5. Process Model 1 (m1)

The process flow form of model 2 (m2) can be seen in the figure 6 When compared with the standard process model then get a rating point value level of 70% so as to obtain the level of capability L (Largely Achieved).

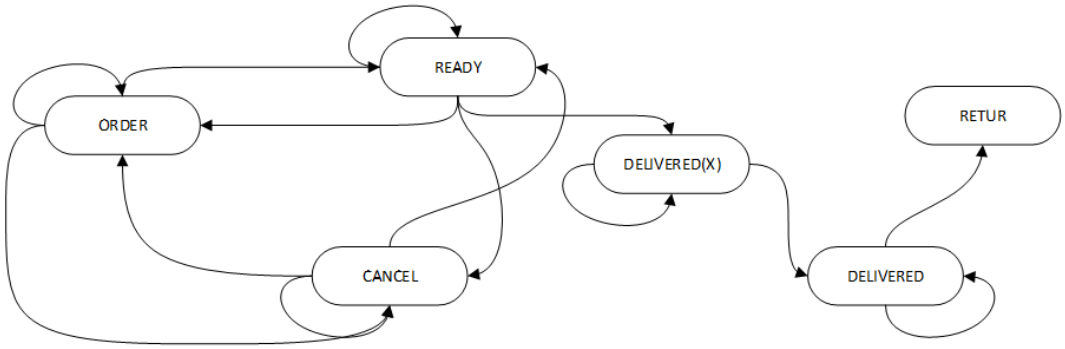

Figure 6. Process Model 2 (m2)

From the result of the attribute rating test on both models, it is found that model 2 has better process matching value than model 1. This is because the second model detects outlier or noise data better, so the trace in the event log Visible is just a common behavior. However, the first model is more representing the actual event log data, it is proved from the *fitness* value obtained by model 1 is bigger than model 2.

4.2 Analysis of Noise Data

The noise data recognized by both models is data that is detected as an outlier or less frequent behavior data. In this study did not specify data suspected of noise as data errors made by the user, because it requires human judgment especially those who understand the actual process flow. Characteristics of noise data itself vary, including missing head, missing tail, and incomplete log. The noise data is recognized from the failed trace replayed by the process model.

From the first model, there are 298 traces that can be replayed, and 39 traces failed to be replayed from a total of 337 trace event log data. Trace recognized as noise in model 1 can be seen in Table 2.

Table 2. Result of noise detection by model 1

Frequency	Trace	Information
11	ORDER	Incomplete
10	READY->DELIVERED(x)->DELIVERED	Missing head
6	CANCEL	Incomplete
2	DELIVERED	Incomplete
1	READY->READY->CANCEL->ORDER->READY->DELIVERED(x)->DELIVERED(x)->DELIVERED(x)->DELIVERED->DELIVERED->DELIVERED	Missing head
1	DELIVERED(x)->DELIVERED	Missing head

From the 39 data traces that failed to be replayed by the first model, there are 31 traces that can be characterized by noise data. The rest needs to be checked again by those who understand the business process to recognize whether the data is noise or not.

From the second model, there are 280 traces that can be replayed, and 57 traces failed to be replayed from a total of 337 trace event log data. Trace recognized as noise in model 2 can be seen in Table 3.

Table 3. Result of noise detection by model 2

Frequency	Trace	Information
11	ORDER	Incomplete
10	ORDER->ORDER	Incomplete duplicated
10	READY->DELIVERED(x)->DELIVERED	Missing head
6	CANCEL	Incomplete
2	DELIVERED	Incomplete
1	ORDER->ORDER->ORDER	Incomplete duplicated
1	ORDER->CANCEL->ORDER	Missing tail
1	DELIVERED(x)->DELIVERED	Missing head

From the 57 data traces that failed to be replayed by the second model, there are 42 traces that can be characterized by noise data. The rest needs to be checked again by those who understand the business process to recognize whether the data is noise or not.

The noise data analysis results show that the second model can recognize more noise data than the first model. This is evidenced in the first model can't recognize the trace of Incomplete duplicated and Missing tail.

4.3 Analysis of Bottleneck process

The analysis was performed on the average time of the longest activity so that it was suspected to cause a bottleneck. Here is the time table 4 for the average of movement activity.

Table 4. Bottleneck between activity

No.	Edge Activity	Frequency	Mean Time (minute)
1	CANCEL – ORDER	56	1445.821
2	CANCEL – DELIVERED(X)	16	700
3	READY – ORDER	53	575.094
4	DELIVERED(X) - DELIVERED	240	565.995
5	ORDER – READY	315	420.396
6	READY – CANCEL	85	367.482
7	ORDER - CANCEL	86	326.186
8	CANCEL – READY	51	318.176
9	READY – DELIVERED(X)	245	309.306

From the calculation of the average time of activity of the activity, it can be concluded that the activity that allows the bottleneck is DELIVERED (X) - DELIVERED, ORDER - READY and READY - DELIVERED (X), due to this activity the amount of frequency is very large event compared to other activities. Then when viewed from the description of the activity in standard operating procedure data that this activity is an important activity, where the customer awaits confirmation of the availability of goods and wait for delivery of goods if the goods are ready.

Conclusion

Based on the result, we state that our method can help auditor to test capabilities of IT processes, by replacing data collection and data validation stage in the real-time process assessment model of COBIT 5 to illustrate the ongoing process represented by event log data. The advantage of this method is the discovery model can be more representative to the original process because it detects only the most frequent behavior and reduce data noise in event log.

The discovery model process of the two proposed models both get high *fitness* value, it means that this method represents the whole process. The second process model can better handle noise data than the first model because it detects more noise and from process attribute rating stage, it shows better compatibility with the standard process model. On enhancement stage, we able to detect which business process that need more time to complete and potentially causing bottlenecks.

This research has some limitations, they are the computation time and resource are very high if the *genetic process mining* handle bigger event log data, so it must be solved by distributed system using parallel computation. And to handle better noise data then required the rule based mechanism. With rule based mechanism then the auditor can be flexible to input rules to detect prohibited process based on their experience.

REFERENCES

[1] W. M. P. v Aalst, K. M. v Hee, J. M. v Werf, and M. Verdonk, *"Auditing 2.0: Using Process Mining to Support Tomorrow's Auditor", Computer,* vol. 43, no. 3, pp. 90-93, *Mar*. 2010.

[2] A. P. Kurniati *and* I. Atastina, *"Implementing Process Mining to Improve COBIT 5 Assessment Program for Managing Operations (Case Study: A University Blog)" ResearchGate*, vol. 72, no. 2, pp. 191-198, *Feb*. 2015.

[3] W. M. P. van der Aalst, *Process Mining – Discovery, Conformance and Enhancement*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[4] W.M.P. van der Aalst, A.K. Alves de Medeiros, and A.J.M.M. Weijters, *"Genetic Process Mining*", Springer-Verlag, Berlin. Vol. 3536, pp 48-69, Jun. 2005.

[5] ISACA. 2012. "*COBIT Process Assessment Model (PAM): Using COBIT 5"*

[6] Suyanto, "*Artifisial Intelligence*", INFORMATIKA, Februari 2011.

[7] I. Tawakkal, "Analisis dan Implementasi Audit Menggunakan *Process Mining* Algoritma *Heuristic Miner* dengan DSS01 COBIT 5", April. 2015.