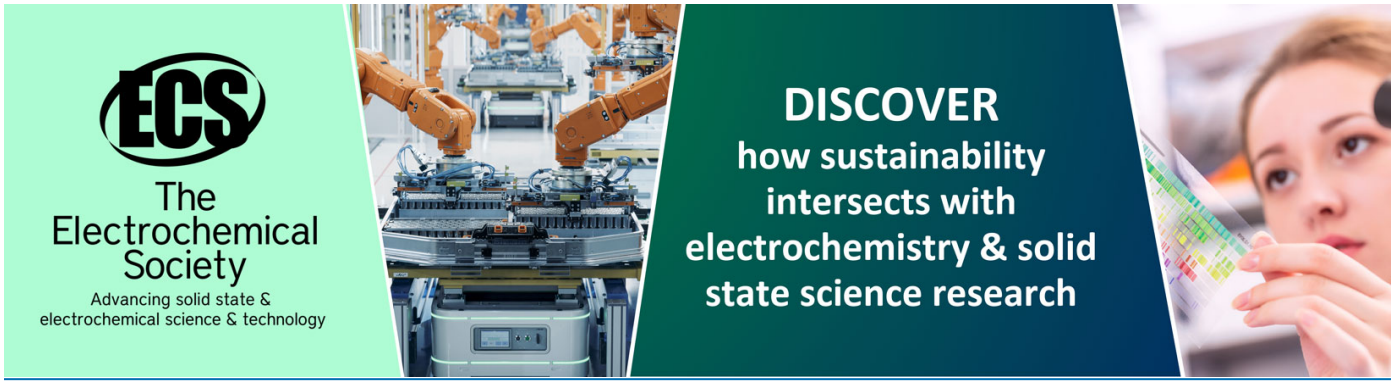
Business process mining from e-commerce event web logs: Conformance checking and bottleneck identification



Business process mining from e-commerce event web logs: Conformance checking and bottleneck identification

M Siek[1*] and R M G Mukti[2]

[1,2] Business Information Systems Program, Information Systems Department,
Faculty of Computing and Media Bina Nusantara University Jakarta, Indonesia 11480

Email: michael.s@binus.edu

Abstract. A range of advanced methods have been formulated and utilized in the efforts of improving the business processes in many enterprises. One impacting enhancement technique is to employ process mining algorithms as modeling and analysis tools in order to provide the actual business performance by digging the event log data and finding the useful information. This paper focuses on the applications of process mining in e-commerce industry. Event log data with timestamps were retrieved and analyzed from the web databases of an e-commerce company and process mining algorithms, like inductive miner and fuzzy miner were executed for generating the actual e-commerce business processes automatically and checking their conformance with the standardized processes as well as to early detecting any bottlenecks and issues in the e-commerce processes. Several e-commerce process issues were considered, such as item procurement, product order and delivery item tracking. The process mining modeling and its statistical results indicate that process mining can provide an efficient and effective tool for modeling and analyzing the e-commerce business processes allowing for real-time process auditing and reengineering.

Keywords: business process modelling, automated model generation, process data analysis, process mining algorithms, real-time process auditing and reengineering, early detection of process bottleneck

1. Introduction

In the recent rapid developments of the advanced technologies and modern business competitions, many companies are inevitably obliged to adapt with some major adjustments in order to stay relevant and competitive to the changing consumer market. A large number of companies do not understand on how they can tackle, implement and leverage these new changes, but they have to attempt to transform their business anyway. In reality, the adoption of the new transformed businesses are not that simple and the people who run the businesses tend to keep their current business processes or in some cases they prefer to perform the business processes manually. The progression of the new process change implementation is often halted when some obstacles related to one of these dimensions, like people, process, method or software, are found. One of advanced technology for business process adjustments and improvements is a so-called process mining [1]. This technique allows for aggressively detecting the business process issues in real-time, yet offering soft adaptation of the new business process implementation through continuous and incremental improvements.

New discoveries and insights inside the actual business processes can be obtained from event log data extracted by using process mining methods [2]. The mining results provide some cause analysis and explanation with a detailed information on why a certain company stands still without any growth. Analyzing process data and modeling business processes are of importance in providing some benefits towards the company competitiveness, such as early detecting the process issues and identification of how to solve those issues. Some interference in business processes, such as inefficiency or ineffectiveness of the processes due to many factors, could disturb the whole of business performance.

The simulation schemes of playing out or playing in of process events are usually equipped in process mining tool for analyzing the process activities and talking process issues if any. The process mining algorithms need event log data as an input data in order to construct business process model and some statistical items for analysis. The event log data contains recorded events or activities with timestamps to specify when a process or activity started and ended [3]. In the example of car manufacturing processes, the casting, stamping, engine assembling, quality controlling and delivery are some of the common processes and activities [4].

The main objective of the research aims at analyzing and implementing process mining techniques in e-commerce industry as one of the most emerging business types that employ much of information system advancements. One of large e-commerce companies was selected as a case study of this research. Some traditional business process modeling was performed to characterize the current the e-commerce business processes. Subsequently, using process mining methods, the researchers automatically generated the e-commerce business processes, checked the conformance between the actual e-commerce business processes with the standardized one and early detected process bottlenecks. Several process issues explored include the restriction of manual processes in some activities (i.e. some sales events required to insert the sales data manually) inducing not up-to-date data or process; and the system integration gaps among entities of salesman, customer and warehouse permitting purchase verification done by customer manually contacting the warehouse people. These process issues are subset of the common problems in e-commerce industry.

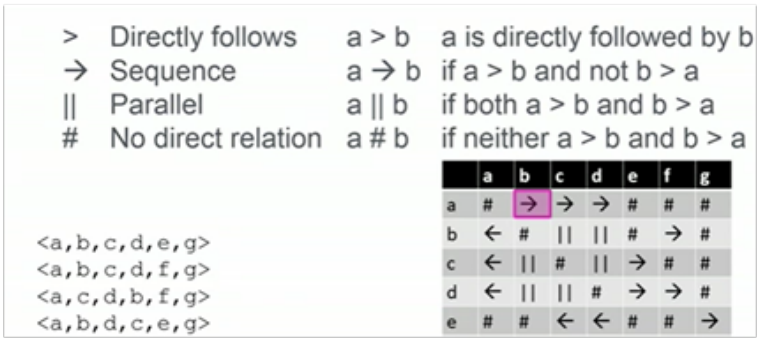
2. Process mining methodologies

2.1. Petri net as universal process representation

In traditional ways of business process analysis and visualization, a Business Process Model (BPM) is commonly used as a business process representation in spite of its weaknesses [5, 6]. Since the good generalization of the resulting process model is crucial in process mining, Petri net notations are mainly utilized. The Petri net enables to represent any business processes as it is more universal. Besides, it is able to visualize the process transitions among activities based on the event log data.

2.2. Event log data

An event log data is a specific database in which all of the events or processes are stored with timestamp information as their footprints. It comprises of three main values, which are Case ID, Activity, and Timestamps; any other data values in an event log data are considered as additional data (i.e. activity cost, person in charge) that can be analyzed further. A company typically has a set of business processes that could consist of a number of process cycles or sub-processes. One process cycle of sub-process is constituted by one scenario in event log data, e.g. a scenario of order process cycle or item return sub-process. A number of the companies probably do not have event logs of their business processes; however they used to have time-stamped records of some activities. The timestamps of the activities can be obtained, for instance, from the sales event held, payment made, product delivered, and so forth.



	a	b	c	d	e	f	g
a	#	→	→	→	#	#	#
b	←	#	||	||	#	→	#
c	←	||	#	||	→	#	#
d	←	||	||	#	→	→	#
e	#	#	←	←	#	#	→

Figure 1. An Example of Event Log Data and Its Notation Depicting the Parallel and Sequential Processes [1]

A scenario comprises of a number of Case IDs or Events that are described by Activity information with timestamps of activity starting and ending date time data. Thus, the event log data with activities and timestamps can present some indications of the time duration required for a particular activity to be accomplished. Therefore, we can figure out how long every activity takes time (duration) and roughly notice that some processes may be executed in parallel or must be done in a sequence.

One example of event log data [1] is shown in Figure 1. One the left bottom of the figure lists four scenarios of processes: <a,b,c,d,e,g>, <a,b,c,d,f,g>, <a,c,d,b,f,g> and <a,b,d,c,e,g>. With brief inspection, it can be understood that a and g are the beginning and the ending of the processes. The activities of b, c and d can be executed in parallel after activity a finishes. The activity d can be followed by either activity e or f, whereas the activities of b and c must continue with the activities of f and e, respectively. On the basis of this basic analysis, we can draw a Petri net diagram [1] to represent these four scenarios of processes as depicted in Figure 2.

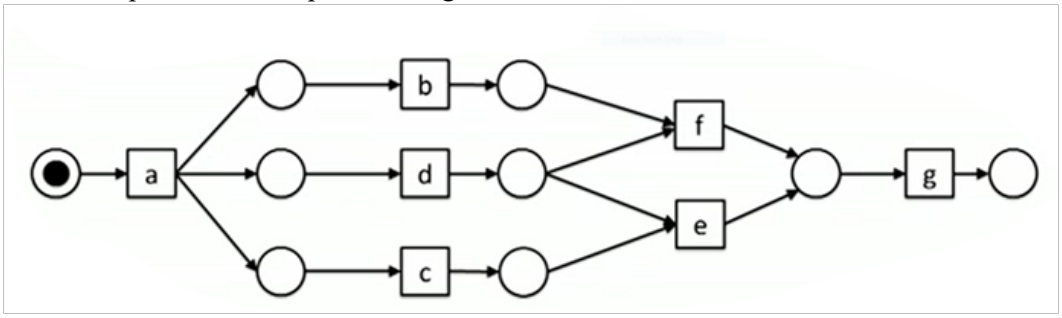


Figure 2. A Petri Net Representation with Some Parallel and Sequential Processes Illustrating the Four Scenarios of the Example

2.3. Process mining algorithms and models

In the view of historical developments of mining algorithms, we have seen the pleasant advancements of artificial intelligence, knowledge-based expert system, machine learning, computational intelligence and data mining [7, 8]. A process mining was firstly initiated as part of developmental efforts thought by the researchers from those larger areas. However, in further progress, the process mining becomes a new area of discipline that was conceptually inspired by the data mining techniques but is vigorously utilized by business process modellers or researchers. The process mining algorithms have significantly been developed and they differ from algorithmic principles in data mining [9].

As a modelling and analysis tool, a process mining provide a visualization of process simulation or play-out of the process mining model generated by the process mining algorithms given an event log data or play-in. A process mining model can be built incrementally as one instance or data sample is

read by a process mining algorithm. Three examples of process mining algorithms [1, 6, 10] include alpha miner, inductive miner and fuzzy miner. The alpha miner basically aspires at building causality reconstruction from a set of event sequences and this algorithm examines the causal relationship between observed activities in sequence. While in inductive miner, the algorithm utilizes recursive procedure to employ divide-and-conquer approach on process discovery by filtering out infrequent processes in every steps of the algorithm. Lastly, the fuzzy miner is specifically used for the large problem of highly unstructured processes by examining significant correlation matrices of the observed processes in the event log data. It allows for more interactive visualization of generated fuzzy model with adjusted level of process abstraction. A fuzzy model [4] generated by fuzzy miner for characterizing car manufacturing processes can be seen in Figure 3.

3. e-Commerce data description and analysis

3.1. Main case study

One large e-commerce company was selected as a case study for this research. The company was found in 2013 as a joint venture between two companies from different countries offering home shopping to the customers. It originally started their home shopping goods and services through local television, commercializing and selling local and imported products at a discounted price for the local consumer. Around 2018, the company launched a browser and mobile application of the home shopping service to the general consumer. The mobile and web app ties with the televised show, but the application often has their own deals and promotions that are separate to the television broadcasting. With respect to the business organizational structure, one company from local country is mainly responsible as the CEO of the joint venture, while the one from overseas manages the operations of the joint company. The software applications utilized for data handling and management, inventory and transaction histories were developed by outsourced software company located in the overseas country.

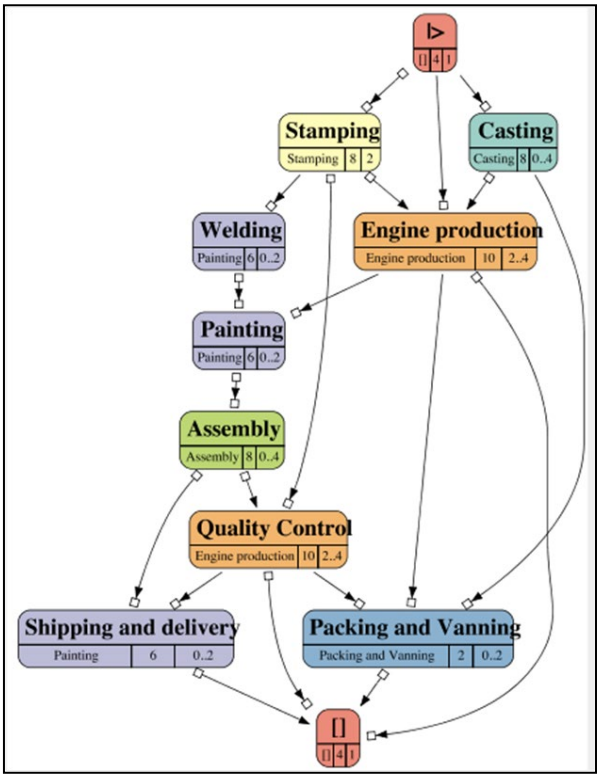


Figure 3. Fuzzy Model for Car Manufacturing Processes Generated by Fuzzy Miner From Event Log Data [4]

The main business processes in this company can be described as follows: (1) the product is commercialized on the television with the hotline number for the customer to call; (2) the customer will call the phone number to make a purchase; (3) the operator on the line will provide necessary information and bank details for payment; (4) the buyer will prepare or do payment depending on the payment method used; (5) the product will be shipped to the customers location; (6) the customer has received the product, which becomes the end process cycle of ordering a product in this e-commerce company. The business processes through online store channel differ slightly from the one through television and hotline channels. Although the process method in delivering a product is the same, the other processes could be different. For instance, after a customer has made an order on the product, an employee from the warehouse should make a call to the buyer to verify the purchase and proceed with the payment method.

3.2. e-Commerce primary process issues

With the ever changing market and ease of access for consumer products, many traditional businesses try to move their strategies forward by making their own online market in order to stay relevant with the current consumer market. Being able to procure and purchase goods or services online, such as buying home appliances or food, and getting a transportation or doctor consultancy is one big trend in e-commerce business strategies. Their services are already online, however some of the companies remain employing manual operations, which do not bring a comprehensive benefit of using their information system. The company as the case study has similar issues as what have been described. In particular, the company has some restrictions that permit for operating certain activities manually and the data and interrelated departments cannot obtain the up-to-date information about a certain activity or process. They still need the transaction data manually for a certain sales event, which is often held and input at the location outside of their IP address domain; hence the transaction on selling a product via this channel needs to be recorded manually to their server.

Moreover, the unrecorded events may be held outside of their main office as a strategy of the company to spread their brand and store to other selling media. Tokopedia, Shopee, Blibli, and Lazada offer this company a place to market their products, which can help the company gain more traction and more income. Subsequently, it lacks of delivery tracking, which the main system cannot provide the information of the current location of the goods and cannot trace them. This results in the purchase cancellation by the customers in the middle stage of delivery process.

3.3. Event log data collection

The event log data for this e-commerce company was acquired and collected from transaction database in several months of the year 2019. The raw data was cleaned up and transformed into event log data structure. Some analyzes against the available event log data were performed by firstly exploiting the inherent business processes of this e-commerce company.

Table 1. A Fragment of Transaction Data in an e-Commerce Company

Order no.	Order type	Order status	Order time	Order media	Order Channel	Category T1	Product code	Product name	Purchase type	Order qty
20191106727106	General Order	Delivery Completed	2019-11-06 11:49:21	TV	Offline Sales	FASHION	203186	CUREISM BRACELET V3	Complete Purchase	2
20191108728772	General Order	Order Cancel	2019-11-08 13:17:54	TV	Outbound	BEAUTY	203758	PAPILUZ HERBAL DIET V1	Consignment Principal	1
20191108728777	General Order	Delivery Completed	2019-11-08 13:25:07	MC	Tokopedia	FASHION	202570	FORSTA MAXIMO WATCH	Consignment Principal	1
20191108728779	General Order	Delivery Completed	2019-11-08 13:26:55	MC	Tokopedia	LIVING GOODS	202588	RE-WASH	Consignment Principal	1
20191108728780	General Order	Delivery Completed	2019-11-08 13:27:20	TV	Outbound	LIVING GOODS	202675	SSENSTORM ALPHA VACUUM CLEANER	Consignment Principal	1
20191108728780	General Order	Delivery Completed	2019-11-08 13:27:20	TV	Outbound	KITCHENWARE	203865	LOCK AND LOCK BISFREE ECO BOTTLE 2 PCS	Complete Purchase	1

4. Results and discussions

4.1. Event log data conversion and analysis

A fragment of transaction data in this e-commerce company was recorded as seen in Table 1. This raw data was required to be cleaned and transformed to as an event log data comprising of Case ID, Activity, and Timestamp fields. Table 2 lists part of event log data converted from the transaction data in the e-commerce company. 'Order No.' was converted to case ID and part of scenario ID in the event log data whereas the time stamps was obtained from the date and time of all activities in the process order cycles, like the date time of the completed events: Order, Payment, Delivery, Receipt. A procedure to data transformation and analysis is the following:

(1) Analyze the data sample and find any data fields that are not relevant to the business processes or not useful for process mining modelling;
(2) Remove unnecessary data values or some anomalies from the process data;
(3) Convert the process data into event log data with proper timestamps
(4) Utilize the event log data in Disco and Rapid Miner Pro-M for process mining modelling and analysis.

This event log data was utilized as the inputs for process mining algorithms to generate the actual e-commerce business processes represented as a Petri net diagram.

Table 2. An Event Log Data Transformed from the Transaction Data in the e-Commerce Company

Order ID	Scenario ID	Case	Activity	Starting datetime	Ending datetime
20190131440451	COD1	3001	Order made	2019/01/31 18:54:37	2019/01/31 18:54:37
20190131440451	COD1	3002	Payment Completed	2019/01/31 18:54:37	2019/01/31 18:54:37
20190131440451	COD1	3003	Shipment	2019/02/01 09:28:59	2019/02/02 19:28:00
20190131440451	COD1	3004	Product Received	2019/02/02 19:28:00	2019/02/02 19:28:00
20190201440566	PG2	4003	Order made	2019/02/01 01:32:23	2019/02/01 01:34:53
20190201440566	PG2	4004	Payment Completed	2019/02/01 01:32:23	2019/02/01 01:34:53
20190201440566	PG2	4005	Shipment	2019/02/01 09:28:59	2019/02/01 21:36:00
20190201440566	PG2	4006	Product Received	2019/02/01 21:36:00	2019/02/01 21:36:00

From the process of data analysis, we discovered some gaps in the data columns, specifically for the payment method of 'cash on delivery'. The buyers with this payment method can cancel their orders, even after the shipment phase of the activity has been carried out. Another finding in process data analysis is the usage of automatic sequential data entries by the system, i.e. column of sequence number. It turns out that the sequence number refers to each individual product purchased with a certain order ID. Table 3 lists the sequential numbers in the second column of the table. This sequential numbering refers to the products 1, 2 and 3 bought by the customer with the same order ID, regardless whether the product has a unique ID or not.

Table 3. Utilization of Sequential Numbering for Identifying the Repeated Orders Issued

20191101722646	001	General Order	Delivery Completed	2019-11-01 00:12:41
20191101722646	002	General Order	Order Cancel	2019-11-01 00:12:41
20191101722646	003	General Order	Delivery Completed	2019-11-01 00:12:41

4.2. Process mining modeling results and analysis

Several process mining tools, such as Disco, Pro-M and Rapidminer Pro-M [10, 11] were utilized for business process analysis and modelling from event log data. A number of tasks were performed from the generation of Petri net process model from event log data, to check the conformance and identify the bottlenecks with some alignment statistic outputs. The researchers exploited three algorithms of alpha, inductive and fuzzy miners to find better generalization and avoid from over-fitting and under-fitting of the resulting models. Each of generated business process models is depicted as Petri net diagram or fuzzy model with frequency information of each process transition for one order process cycle. The frequency information is depicted as an arrow line and frequency number in between two events. The thicker the arrow line, the more frequent the process leading to the next process. With the features of playing-in and play-out of the models, we can simulate the actual business processes, check their conformance, and identify the bottlenecks as well as provide accurate solutions.

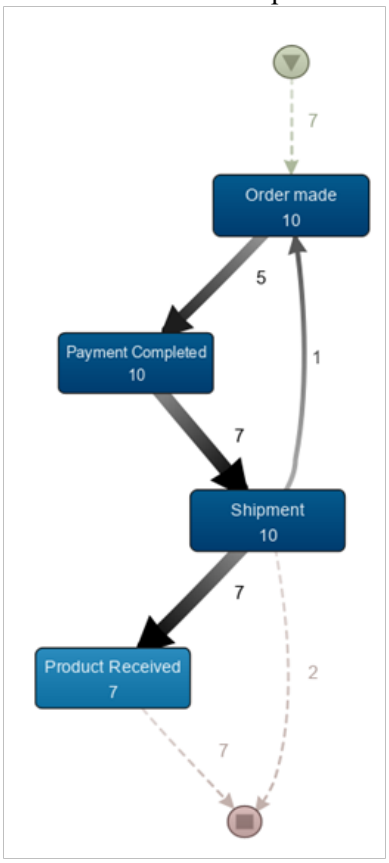


Figure 4. Business Process Model Generation from the e-Commerce Event Log Data using Inductive Miner Algorithm

Through mining the event log data, the inductive miner was able to generate the business processes automatically, as seen in Figure 4. This resulting process model closely represents the actual business processes in this e-commerce company. The process modelling shows the frequencies of two consecutive events occurred depicted as arrow thickness and numbers. The numbers in the activity boxes indicates the frequency of each activity executed. It can also be inferred based on the scenarios that there were some rare occurrences for completed product received. The reason for having this case is due to the fact that the buyers are able to cancel their orders as soon as the products arrived at the customer's hands.

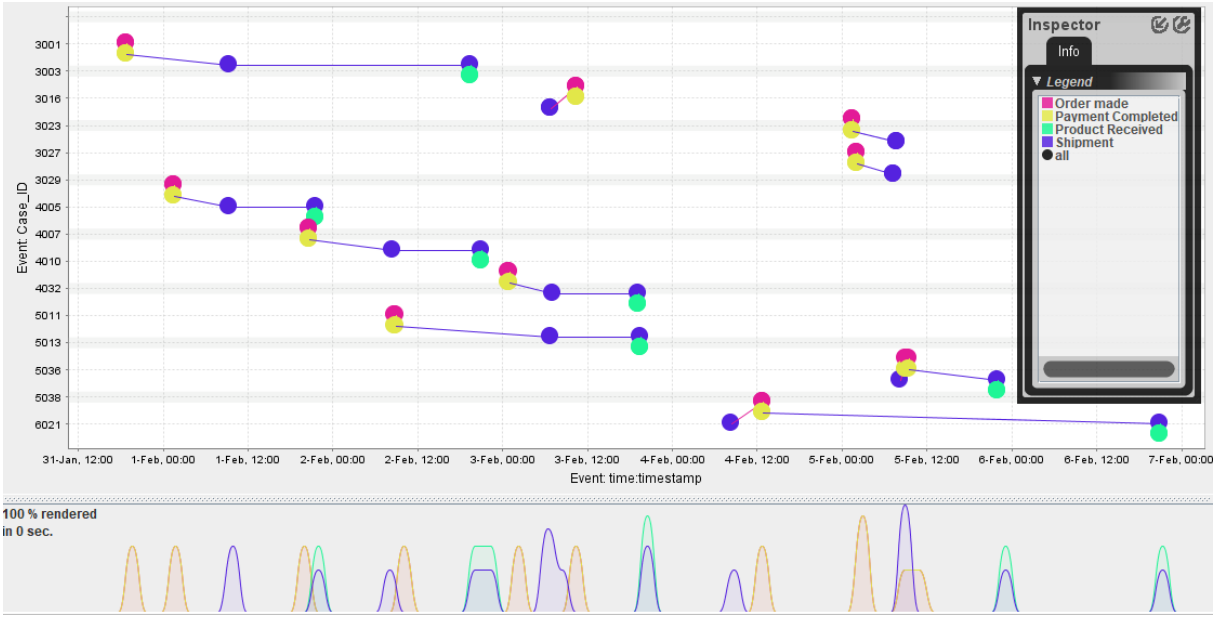


Figure 5. Dotted Chart with Event Names of Case ID Versus Timestamp Depicting Some Variations on Time Duration and Frequency on e-Commerce Order Transaction Processes.

One important chart to be analyzed is a dotted chart that has event names of Case ID versus Timestamps, as illustrated in Figure 5. This chart supplies some indications on checking how often a particular activity in e-commerce business processes requires to be performed in regards to all scenarios. In this dotted chart, we can see the list of events that have occurred in both of the activity and event log with the timestamps placed for handling the event separation based on its frequency and occurrence. It becomes apparent in this dotted chart that the least frequent event executed in the process is Product Received. It illustrates the time duration between two consecutive activities describing the whole order process cycle in the e-commerce company, from a starting process of product ordering through various selling channels to the ending process of product received and/or with payment in the case of cash in delivery. Furthermore, it can be seen that some of the process scenarios end at the event of Shipment. This is due to the fact that the customers cancelled their product orders in the middle process of delivery and this is the case of using cash on delivery payment method, thus the event of Product Received was not recorded in the event log data.

Table 4. Statistical Evaluation Measures to Indicate the Process Alignments

	Average	Max.	Min.	Std. Deviation	#Case Value of 1.00
Raw Fitness Cost	1.70	3	1	0.82	5
Calculation Time (ms)	16	16	16	0.00	0
Num. States	21	25	18	3.27	0
Move-Model Fitness	1.00	1	1	0.00	10
Trace Fitness	0.79	0.89	0.67	0.10	0
Move-Log Fitness	0.76	0.88	0.62	0.12	0
Trace Length	7.40	8	6	0.97	0
Queued States	44.10	46	42	1.52	0

The alignment frequencies of a certain process conforming to the standard predefined process can be indicated by statistical evaluation measures, such as move-log fitness, raw fitness cost, trace fitness and move-model fitness, as listed in Table 4. The move-model fitness of 1, trace fitness of 0.79 and move-log fitness 0.76 indicate reliable conformance or alignment, whereas the raw fitness cost of 1.70 do not provide good indicator of conformance. The latter is due to the fact that the number of cases for every scenario in event log data is statistically not sufficient for the process mining to determine the conformance.

5. Conclusions

Despite more significant numbers of businesses transforming to e-commerce, the leverages of process mining in the e-commerce industry are quite promising and encouraging. Process mining techniques from event log data have offered some advances on automated generation of business process model, conformance checking between the operating business processes and the standardized ones, and early bottleneck identification with some alternative accurate solutions. A number of e-commerce process issues such as gaps in system integration, less restriction for manual processes, manual payment verification and order cancellation for cash in delivery payment methods, can be immediately detected in the process mining modeling. The usage of process mining analysis and modeling in e-commerce industry permits for soft-approach on improving their business processes through continuous and incremental advancements, which could lead to the capabilities for automated real-time process auditing and reengineering in e-commerce industry.

Acknowledgement

This work is supported by Research and Technology Transfer Office, Bina Nusantara University as a part of Bina Nusantara University's International Research Grant entitled "Big Data Analytics and Computational Intelligence for Automated Decision Making" with contract number: No.026/VR.RTT/IV/2020 and contract date: 6 April 2020.

References

[1] van der Aalst W M 2016 *Process Mining. Data science in action*: Springer-Verlag Berlin.

[2] van der Aalst W M 2011 Process mining: Discovery, conformance and enhancement of business processes (Berlin: Springer-Verlag).

[3] Ghasemi M and Amyot D 2019 From event logs to goals: a systematic literature review of goal-oriented process mining *Requirements Engineering* p 1-27.

[4] Siek M and Mukti R G M 2019 Process mining with applications to automotive industry *International Conference on Multidisciplinary Research*, Bogor, Indonesia.

[5] van der Aalst W M, Adriansyah A, de Medeiros A K A, Arcieri F, Baier T, Blickle T*, et al.* 2011 Process mining manifesto *International Conference on Business Process Management* p 169-194.

[6] Leemans S J, Fahland D, and van der Aalst W M 2013 Discovering block-structured process models from event logs-a constructive approach *International conference on applications and theory of Petri nets and concurrency* p 311-329.

[7] Solomatine D P and Siek M B 2006 Modular learning models in forecasting natural phenomena *Neural networks* **19** p 215-224.

[8] Siek M 2011 *Predicting storm surges: Chaos, computational intelligence, data assimilation and ensembles* (CRC Press).

[9] van der Aalst W M, Reijers H A, Weijters A J, van Dongen B F, de Medeiros A A, Song M*, et al.* 2007 Business process mining: An industrial application *Information Systems* **32** p 713-732.

[10] Mans R, van der Aalst W M and Verbeek H 2014 Supporting process mining workflows with RapidProM *BPM (Demos)* p 56.

[11] van der Aalst W M, Bolt A and van Zelst S J 2017 RapidProM: Mine your processes and not just your data (Cornell University) *arXiv preprint* https://arxiv.org/abs/1703.03740