khazanah informatika

Jurnal Ilmu Komputer dan Informatika

The Design of Exploratory and Preprocessing of Event Log Data in Online Learning Activities Based on Moodle LMS for Process Mining

Demaspira Aulia*, Indra Waspada
Departemen Ilmu Komputer/Informatika, Fakultas Sains dan Matematika
Universitas Diponegoro
Semarang
*demaspira@student.undip.ac.id

Abstract-Process Mining is one of the sub-studies of Data Mining that focuses on the events of a system. An area that benefits from process mining is education, especially online learning. This study used Moodle as a platform to provide online event activity log data in online learning. Moodle-based process mining requires several stages that are not easily understood directly by teachers. As a solution, some efforts are needed to integrate Moodle with process mining. This study built an application that could contribute to the Preprocessing and Exploratory Data Analysis (EDA) stages of Moodle event log data – as an important part of the process mining stage. Preprocessing was implemented by using the simple heuristic filtering method, while EDA was employed through visualization using flow control and dotted charts. Eventually, the application built in this study successfully performed preprocessing in Moodle event log data and could display the results visually, as a tool of control flow analysis and dotted chart analysis.

Keywords: exploratory data analysis, Moodle, process mining

1. Introduction

The use of Data Mining (DM) has been often used in several fields. One of the fields is the educational field. The goal of implementing the DM process is to find interesting patterns from large data [1]. The use of DM in the educational field is also known as Educational Data Mining (EDM). EDM has two types of objectives which are to improve the learning process and to gain an understanding towards the learning phenomenon [2].

In addition to using DM, process mining has been recognized to be used in various fields, especially in business field. Process mining is one of the new research disciplines between machine learning and data mining using process modelling and analysis. The objectives of process mining included finding, monitoring, and improving the processes that occur by taking existing knowledge from the event log obtained from the system [3]. From these objectives, the process mining can also be applied to other fields that are in education – known as Educational process mining.

EDM employs an event log that is obtained to find, monitor, and improve the educational process that is applied [4]. In the implementation of EDM, the perspective that is the most often used is control flow perspective – focuses on the sequence of existing activities [5]. By using control flow perspective, we could see the application of the patterns in teaching and learning process. However, the control flow used in this study is still limited to EDA and does not cover the process mining part. Another perspective that is frequently applied in educational process mining is a performance perspective that is depicted in the form of a dotted chart. The dotted chart is used to observe the flow of events occurs based on the event log [6].

One of the implementations of online learning can be applied to the Learning Management System (LMS). One of LMSs that is often used is Moodle. Moodle is designed to help teachers who try to make quality online learning. Moodle is used in various universities, schools, and companies throughout the world. Moodle helps teachers to arrange lessons in various ways and integrate lessons with collaborative activities [7]. Moodle saves data in a MySQL database – storing an event log that occurs on the system is done into several tables in the database [8]. By accessing the database, the event log can be obtained from the learning process.

A previous study conducted by Slaninova *et al.* [8] uses data from event log Moodle to analyze students' behavior in Moodle; group them based on behavioral similarities between students; and also visualize the relationship between students and the groups. Besides, Bogarin *et al.* [9] used clustering algorithms on students' interaction data

in Moodle to improve the existing process mining models. Another study conducted by Bogarin *et al.* [10] also use clustering and process mining to find navigation paths of students in Moodle. A research conducted by Juhanak *et al.* [11] analysed the patterns of students' behaviour and interaction – which were different in each quiz attempt in LMS. Therefore, by using process mining, the sequence of activities carried out by students at the time of quizzes can be obtained.

Exploratory Data Analysis (EDA), according to Willems [12] is used to answer questions and business assumptions as well as to make hypotheses for further analysis. On top of that, the function of EDA is to prepare data for Modelling. It is known that by having good knowledge regarding the data used could provide the answers needed or could build intuition to interpret the results of modelling that were conducted. Stages in EDA include describing the data to be used, taking samples from the data, overcoming problems in the data, understanding the features in the data, and understanding the pattern of the data.

The role of EDA and preprocessing as the initial stages of using process mining is considered important [13][14], especially because there are several problems in the implementation of process mining including the poor quality data and other issues related to data quality which had not been frequently raised by researchers [15]. The learning process that became a focus in this research was the process of quizzes for students or participants in Moodle. Thus, we could find out the flow and order of the quiz occurred in Moodle.

Although there have been several studies that use process mining in education, especially those using Moodle, there was no integration between process mining and Moodle which results in the difficulty for teachers and researchers to obtain data and information from process mining results on the Moodle data.

The contribution given to this research is in the form of application design as an initial stage of the integration of process mining on Moodle. The integration conducted in this study was still simulated and not directly integrated into Moodle. Additionally, the data employed still referred the exported data from Moodle. In this application, EDA was performed on the event log data of Moodle, which was done as the initial stage before the conducting process mining. The stages in building the application commenced from data identification, requirement analysis and definition, case studies identification, application development result, application testing and experiments with specified scenarios to prove that the application can be applied to original data from Moodle.

2. Method

a. Application Architecture

The application developed in this study is a part of the system that was built to integrate Moodle with process mining. The aim of this application was to carry out the EDA and preprocessing in the event log prior to the implementation of process mining. The data used was obtained by downloading the event log through Moodle and uploading it to the application.

This application is divided into two parts – including web server and client. The client is used as an intermediary between users and web servers for preprocessing and EDA. Moreover, the client also has a function to display graphs based on data sent by the webserver. The web server acts as a data processor and handles all other processes that are being executed.

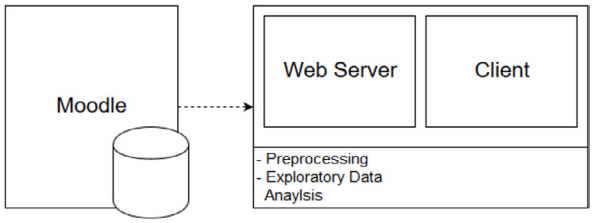


Figure 1. Application Architecture

b. Data Identification

The data used in this application were online learning data obtained in the form of an event log from Moodle. The data used were in the form of CSV files. The number of attributes needed to use this application did not have a maximum limit but had at least three main attributes as mentioned by Aalst [3]: case id, event, and timestamp. The case id and event can change according to the context [15]. By this, this study has two types of scenarios using different case id and event. An explanation of the attributes in the data can be seen in Table 1.

Table 1. Description of the event log data attributes from Moodle

Attributes	Descriptions
Time	Time marker when the event occurs
User full name	Complete name of the event user
Affected user	Name of the event target
Component	Marker of event types in general
Event context	Marker of the event conducted
Event name	Name of the event conducted
Description	Description of the event
Origin	Origin of the event
IP address	IP address of the event users

From the data, there was a problem with the Time attribute in which the value on that attribute could not be directly used as timestamp since the format was not related to ISO standard. Consequently, it made the application difficult to use the value of the Time. Therefore, in the preprocessing stage of the application, there was an option offered to change the format of

Time attribute to conform to ISO standard.

c. Requirements Analysis and Definition

To achieve the goals and solve the problems that is found in data identification, it is necessary to identify the required features that were used in this application.

Table 2. Results of features identification of ProM

No	Name of Feature	Description
1	Import data in csv format	Receiving the Moodle event log data in csv format
2	Data column setting	Choosing three required main columns including case id, event, and timestamp
3	Summary of data statistic	Displaying the information in the form of the statistic of start event, end event, and whole event of preprocessed data
4	Simple heuristic filtering	Conduct filtering using simple heuristic filtering
5	Control flow analysis	Visualization using *control flow perspective*
6	Dotted chart analysis	Visualization using the performance's perspective of the *dotted chart*. Visualization given consisted of two types of *dotted chart* including *dotted chart* based on absolute time and relative time.

Table 3. Additional features based on problems in data

No	Name of feature	Description
1	Preprocessing data	Preprocessing on data – consisting of time format conversion, alias and initial filter naming, column combination, quiz attempt calculation, column deletion, and data column setting.

The first step was analyzing the features of the existing process mining tool called ProM. ProM is an open source framework that supports various forms of process mining techniques. The version of ProM analyzed in this study was ProM Lite version 1.2. Based on the analysis results, the features of ProM that can be implemented in the application is informed in Table 2.

Table 4. SRS lists of the application

No	ID SRS	Description
1	SRS-F-01	Receive event log data in csv format
3	SRS-F-02	Preprocess data
4	SRS-F-03	Display summary of data statistic
5	SRS-F-04	Analyse control flow
6	SRS-F-05	Analyse dotted chart

Based on the results of the feature identification in Table 2 and Table 3, the features to be implemented will be used as a reference to determine the functional requirements of the application. The functional requirements are represented in the form of Software Requirement Specification (SRS) in which the SRS list of applications can be seen in Table 4.

The requirements of preprocessing data (SRS-F-02) includes the following parts:

1. Time format conversion
2. Data alias presentation
3. Column combination
4. Quiz attempt calculation
5. Column deletion
6. Data column setting
7. Filtering using simple heuristic filtering.

d. Case Study

The data used was lecture data in Basic System course in semester 1 of 2018/2019 Academic Year at the Department of Informatics, Diponegoro University. The lecture utilized the Moodle platform used by Undip Informatics, which could be accessed at https://ioclass.if.undip.ac.id/.

The data included event log from all activities conducted by lecturers and students. However, this study was focused on quiz work at IOClass only. The data used contained data starting from the beginning of lectures to the end of semester exams. The data has 9 columns and 178920 rows. The examples of the scores from the data used is informed in Table 5.

Table 5. The example of the content of event log data

Time	User full name	Affected user	Event Context	Component	Event Name	Description	Origin	IP address
24/09/2018, 15:45	SDAW	SDAW	Quiz: Quiz 5: Boole Algebra and Combinational Logic	Quiz	Quiz Attempt Viewed	The user with id '845' has viewed the attempt with id '9206' belonging to the user with id '845' for the quiz	web	182.1.68.53
24/09/2018, 15:44	SDAW	SDAW	Quiz: Quiz 5: Boole Algebra and Combinational Logic	Quiz	Quiz Attempt Viewed	The user with id '845' has viewed the attempt with id '9206' belonging to the user with id '845' for the quiz	web	182.1.68.53

Time	User full name	Affected user	Event Context	Component	Event Name	Description	Origin	IP address
24/09/2018, 15:44	SDAW	SDAW	Quiz: Quiz 5: Boole Algebra and Combinational Logic	Quiz	Quiz Attempt Viewed	The user with id '845' has viewed the attempt with id '9206' belonging to the user with id '845' for the quiz	web	182.1.68.53
24/09/2018, 15:43	SDAW	SDAW	Quiz: Quiz 5: Boole Algebra and Combinational Logic	Quiz	Quiz Attempt Viewed	The user with id '845' has viewed the attempt with id '9206' belonging to the user with id '845' for the quiz	web	182.1.68.53

The quiz setting applied to "Basic Systems" course was that each student was allowed to attempt many times on each quiz – yet there was a specified time limit. In case a quiz attempt reaches the time limit, it was automatically collected into the system. By applying this setting, it was possible for students to conduct the same quiz repeatedly. Also, the final quiz scores taken were the highest scores from the quiz's attempts conducted. This setting applied to all quizzes, in addition to the midterm and final semester exams.

e. Experiment Scenarios

The experiment used Moodle event log data from "Basic Systems" course. The experiment that was carried out consisted of two scenarios. The first scenario produced data that contained case id where participants with the whole quiz data as event and producing data containing case id where each quiz attempted by each participant and quiz data separated considered as event.

Table 6. Attribute mapping on scenario 1

Attribute	Form	Example
case id	participantName	Demaspira Aulia
event	eventName	Quiz attempt started

Table 7. Attribute mapping on scenario 2

Attribute	Form	Example
case id	participantName_ noAttempt	Demaspira Aulia_12
event	quizName_ eventName	Quiz 1_Quiz attempt started

The objectives of these two scenarios were to find possible quiz patterns on the Moodle platform and to analyse whether the patterns are in accordance with the actual events. Other objectives of both scenarios were to observe the duration the quizzes attempted by students and to find patterns that are considered abnormal.

The mapping of the two scenarios is informed in Table 6 and Table 7. The visible difference between the two scenarios is that in scenario 1, general data without any specific quiz details was the only visible result in scenario 1. Whereas, in scenario 2, specific data for each experiment attempted on each quiz was clearly observed.

3. Results

a. The Result of Application Development

The results of the application design explained in the previous chapter were then implemented into a web-based application using the Python and Typescript programming languages. Both programming languages are chosen based on their performance and ease of application on the web platform. Python is used to implement the data processing functions that contained in the application. Whereas, Typescript is utilized to display the data that was processed using Python into an easier and more understandable form.

The needs for receiving data in CSV format (SRS-F-01) was implemented into the web page which was used to upload the event log data employed. The data was stored on a storage server later on and was used as a reference for the further process of initial data. The interface – as the result of SRS-F-01 – can be seen in Figure 2.

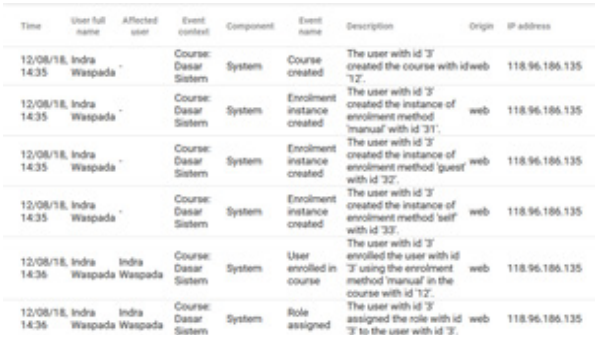
The needs of data preprocessing (SRS-F-02) are implemented into a web page consisting of six parts including time conversion, data alias assignment, attempt quiz calculation, merging two columns, columns deletion, and three main columns selection that was used as case id, event, and timestamp. The time conversion section has a function to convert the time format from a column into ISO format, data alias assignment has a function to pre-filter the values from a column that are not required. In addition, another name or abbreviated name was given to the values of a column to facilitate the user in reading the data. Subsequently, the function of the quiz attempt calculation was to calculate the number of quiz attempts conducted by participants.

This is intended to distinguish the same quiz attempt conducted by certain participants. Furthermore, merging two columns benefited to merge the scores from two different columns. Also, column deletion was used to eliminate unneeded columns and the selection of three main columns was intended to select the case id, event, and timestamp of the data used. Hence, other features of this application could be applied as well. In this case, filtering was also conducted by employing simple heuristic filtering – to select the values used at the start event, end event, and all events. This is intended to eliminate data

that may include outliers from existing data. The interface – result of SRS-F-02 – is illustrated in Figure 3 and the interface of simple heuristic filtering is shown in Figure 4.

The need of displaying the statistics of data (SRS-F-03) was used to show statistical information of data that had already been subjected to preprocessing and/or filtering. The information displayed included the number of existing cases, the total number of existing events, the number of classes of events, the frequency of start events, end events, and the whole events (all events) of existing cases. The interface of SRS-F-03 is depicted in Figure 5.

The need of control flow analysis (SRS-F-04) and dotted chart analysis (SRS-F-05) was used to provide an overview of the existing event flow using a control flow chart and dotted chart. The dotted chart consisted of two types including dotted chart using absolute time that was used to show how and when each case begins and how the flow occurs. Moreover, dotted chart using relative time benefited to observe the performance of the time of each existing case. The interface implementation of SRS-F-04 and SRS-F-05 are shown in Figure 6 and Figure 7, respectively.



Time	User full name	Affected user	Event context	Component	Event name	Description	Origin	IP address
12/08/18, 14:35	Indra Waspada		Course: Dasar Sistem	System	Course created	The user with id '3' created the course with id '12'.	web	118.96.186.135
12/08/18, 14:35	Indra Waspada		Course: Dasar Sistem	System	Enrolment instance created	The user with id '3' created the instance of enrolment method 'manual' with id '31'.	web	118.96.186.135
12/08/18, 14:35	Indra Waspada		Course: Dasar Sistem	System	Enrolment instance created	The user with id '3' created the instance of enrolment method 'guest' with id '32'.	web	118.96.186.135
12/08/18, 14:35	Indra Waspada		Course: Dasar Sistem	System	Enrolment instance created	The user with id '3' created the instance of enrolment method 'self' with id '33'.	web	118.96.186.135
12/08/18, 14:36	Indra Waspada	Indra Waspada	Course: Dasar Sistem	System	User enrolled in course	The user with id '3' enrolled the user with id '3' using the enrolment method 'manual' in the course with id '12'.	web	118.96.186.135
12/08/18, 14:36	Indra Waspada	Indra Waspada	Course: Dasar Sistem	System	Role assigned	The user with id '3' assigned the role with id '3' to the user with id '3'.	web	118.96.186.135

Figure 2. Result interface of SRS-F-01

case_id	task	timestamp	User full name	Event context	Component	Event name	n_attempt
147_1	Quiz: quiz 1 - pengenalan dunia digital_qas	2018-08-21 07:00:00	147	Quiz: quiz 1 - pengenalan dunia digital	Quiz	qas	1
147_1	Quiz: quiz 1 - pengenalan dunia digital_qav	2018-08-21 07:00:00	147	Quiz: quiz 1 - pengenalan dunia digital	Quiz	qav	1
147_1	Quiz: quiz 1 - pengenalan dunia digital_qav	2018-08-21 07:01:00	147	Quiz: quiz 1 - pengenalan dunia digital	Quiz	qav	1
147_1	Quiz: quiz 1 - pengenalan dunia digital_qav	2018-08-21 07:01:00	147	Quiz: quiz 1 - pengenalan dunia digital	Quiz	qav	1
147_1	Quiz: quiz 1 - pengenalan dunia digital_qav	2018-08-21 07:01:00	147	Quiz: quiz 1 - pengenalan dunia digital	Quiz	qav	1
147_1	Quiz: quiz 1 - pengenalan dunia digital_qav	2018-08-21 07:02:00	147	Quiz: quiz 1 - pengenalan dunia digital	Quiz	qav	1

Figure 3. Result interface of SRS-F-02

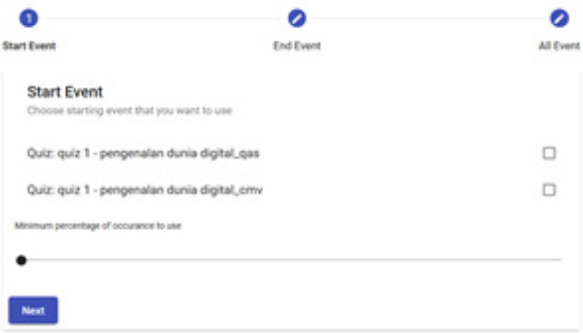


Figure 4. Filtering start event interface in SRS-F-02

Figure 5. Statistic interface of start event of SRS-F-03

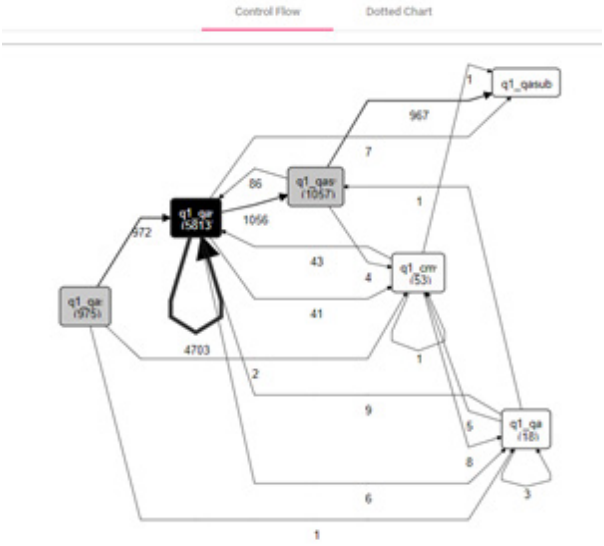


Figure 6. Interface implementation of SRS-F-04 (control flow analysis)

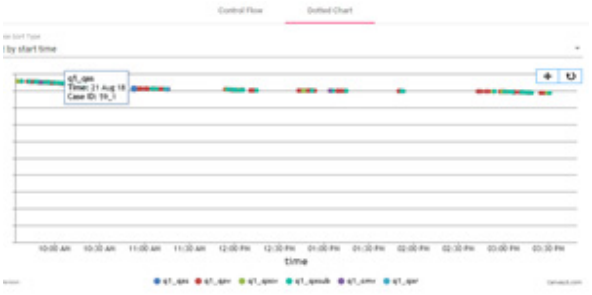


Figure 7. Interface implementation of SRS-F-05 (dotted chart analysis)

b. Experiment

The experiment was conducted using original Moodle data taken from online lecture data from IOClass Undip Informatics "Basic System" course. The experiment was conducted to prove that this application was able to run in accordance with predetermined needs using real data. Table 8 informs the steps conducted, the process worked on, and the features used in each trial scenario.

Table 8. Table of experiment steps of scenario 1 and 2

Scenario	*Preprocessing*	Used features	*Simple Heuristic Filtering*
Scenario 1	1. Converting time in the time column 2. Filtering the component column by simply filling in the Quiz value 3. Filtering user full name column by removing grades from a lecturer, assistant, or administrator 4. Giving alias to the value in the event name column 5. Giving alias in the form of a number to value in the user full name column 6. Deleting affected user column and the IP address 7. Selecting user full name column as the case id, event name column as event and time as the timestamp	- Time format conversion - Giving alias names - Column deletion - The setting of the data column	No filtering conducted in simple heuristic filtering
Scenario 2	1. Converting time in the time column 2. Filtering the component column by simply filling in the quiz value 3. Filtering user full name column by removing grades from the lecturer, assistant, and administrator 4. Giving alias to the value in the event name column 5. Giving alias to value in the event context column and conducting filtering process by only taking values of certain quiz 6. Giving aliases in the form of sequential numbers in the value in the User full name column 7. Calculating the number of attempts using user full name column as base column, event name column as count column, selecting quiz attempt started as start event, quiz attempt submitted as end event, and generating the n_attempt column 8. Merging the event name column with Event context using underscore delimiter (_) and then saving it with the Event name 9. Merging the user full name column with n_attempt column using underscore delimiter (_) and then saving it with case id name 10. Deleting affected user and the IP address columns 11. Selecting case id column as case id, event column as event, and time column as the timestamp	- Time format conversion - Giving alias names - Calculation of quiz attempts - Columns merging - Column deletion - Column setting - Simple heuristic filtering	The minimum occurrence value of the event that was used in start event and end event, was 90%. Also, for all event, the entire events were used

Table 9. Table of Aliases event name column

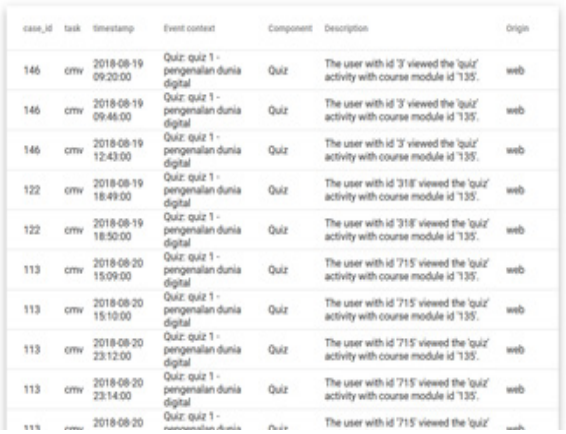
No	Initial score	Alias	Description
1	*Quiz attempt started*	Qas	Quiz is started
2	*Course module viewed*	Cmv	Participants view the module
3	*Quiz attempt viewed*	Qav	Working on the quiz
4	*Quiz attempt submitted*	Qasub	Quiz results are submitted
5	*Quiz attempt summary viewed*	Qasv	The summary of the quiz before it is submitted
6	*Quiz attempt abandoned*	Qaban	Quiz attempt aborted
7	*Quiz attempt reviewed*	Qar	Review the results of the quiz

Giving alias names in certain columns is intended to ease data reading. For the User full name column, the alias name is intended to maintain the confidentiality of quiz participants' identity. Details of the aliases for the event name column and event context column are shown in table 9 and table 10, respectively.

Table 10. Table of aliases in event context column

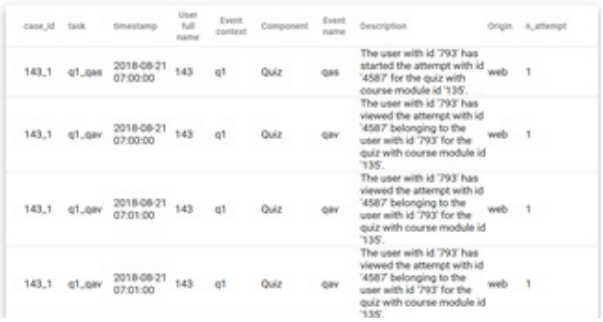
No	Initial Score	Alias
1	Quiz: quiz 1 – introduction to digital world	q1
2	Quiz: Quiz 2 : Digital system	q2
3	Quiz: Quiz 3: number conversion	q3
4	Quiz: Quiz 4: Basic logic	q4
5	Quiz: Quiz 5: Boole Algebra and combinational logic	q5
6	Quiz: Quiz 6: Comparator	q6
7	Quiz: Quiz 7: Combinational logic adder substractor	q7
8	Quiz: quiz 8: mux & ndcoder	q8
9	Quiz: Quiz 9: sequencial logic - FF	q9
10	Quiz: Kuis 10: Register, Counter, ROM	q10
11	Quiz: Responsi 1	qr1
12	Quiz: Responsi 2	qr2
13	Quiz: UTS 1	quts
14	Quiz: UTS 2	quts

It is observed in Table 16 that each Quiz responsi and UTS Quiz are given the same alias name – Quiz responsi is given the alias name qr1 and qr2 while UTS Quiz is given the alias name quts.



case_id	task	timestamp	Event context	Component	Description	Origin
146	cmv	2018-08-19 09:20:00	Quiz: quiz 1 - pengenalan dunia digital	Quiz	The user with id '3' viewed the 'quiz' activity with course module id '135'.	web
146	cmv	2018-08-19 09:46:00	Quiz: quiz 1 - pengenalan dunia digital	Quiz	The user with id '3' viewed the 'quiz' activity with course module id '135'.	web
146	cmv	2018-08-19 12:43:00	Quiz: quiz 1 - pengenalan dunia digital	Quiz	The user with id '3' viewed the 'quiz' activity with course module id '135'.	web
122	cmv	2018-08-19 18:49:00	Quiz: quiz 1 - pengenalan dunia digital	Quiz	The user with id '318' viewed the 'quiz' activity with course module id '135'.	web
122	cmv	2018-08-19 18:50:00	Quiz: quiz 1 - pengenalan dunia digital	Quiz	The user with id '318' viewed the 'quiz' activity with course module id '135'.	web
113	cmv	2018-08-20 15:09:00	Quiz: quiz 1 - pengenalan dunia digital	Quiz	The user with id '715' viewed the 'quiz' activity with course module id '135'.	web
113	cmv	2018-08-20 15:10:00	Quiz: quiz 1 - pengenalan dunia digital	Quiz	The user with id '715' viewed the 'quiz' activity with course module id '135'.	web
113	cmv	2018-08-20 23:12:00	Quiz: quiz 1 - pengenalan dunia digital	Quiz	The user with id '715' viewed the 'quiz' activity with course module id '135'.	web
113	cmv	2018-08-20 23:14:00	Quiz: quiz 1 - pengenalan dunia digital	Quiz	The user with id '715' viewed the 'quiz' activity with course module id '135'.	web
		2018-08-20	Quiz: quiz 1 -		The user with id '715' viewed the 'quiz'	

Figure 8. Preprocessing results of scenario 1



case_id	task	timestamp	User full name	Event context	Component	Event name	Description	Origin	n_attempt
143_1	q1_qas	2018-08-21 07:00:00	143	q1	Quiz	qas	The user with id '793' has started the attempt with id '4587' for the quiz with course module id '135'.	web	1
143_1	q1_qav	2018-08-21 07:00:00	143	q1	Quiz	qav	The user with id '793' has viewed the attempt with id '4587' belonging to the user with id '793' for the quiz with course module id '135'.	web	1
143_1	q1_qav	2018-08-21 07:01:00	143	q1	Quiz	qav	The user with id '793' has viewed the attempt with id '4587' belonging to the user with id '793' for the quiz with course module id '135'.	web	1
143_1	q1_qav	2018-08-21 07:01:00	143	q1	Quiz	qav	The user with id '793' has viewed the attempt with id '4587' belonging to the user with id '793' for the quiz with course module id '135'.	web	1

Figure 9. preprocessing results of scenario 2

The resulted data of preprocessing stage is shown in figure 8 and figure 9. It is observed that the difference between the two scenarios lies on the case id and the event used.

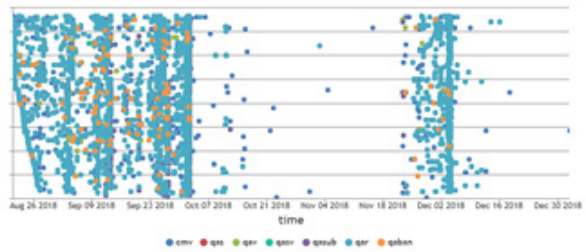


Figure 10. Dotted chart using absolute time in scenario 1

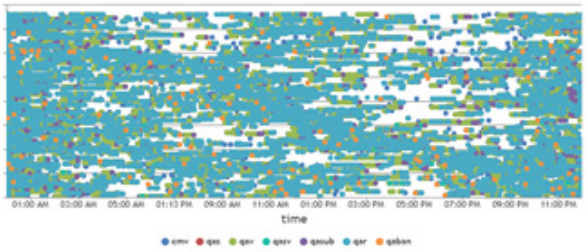


Figure 11. Dotted chart using relative time in scenario 1

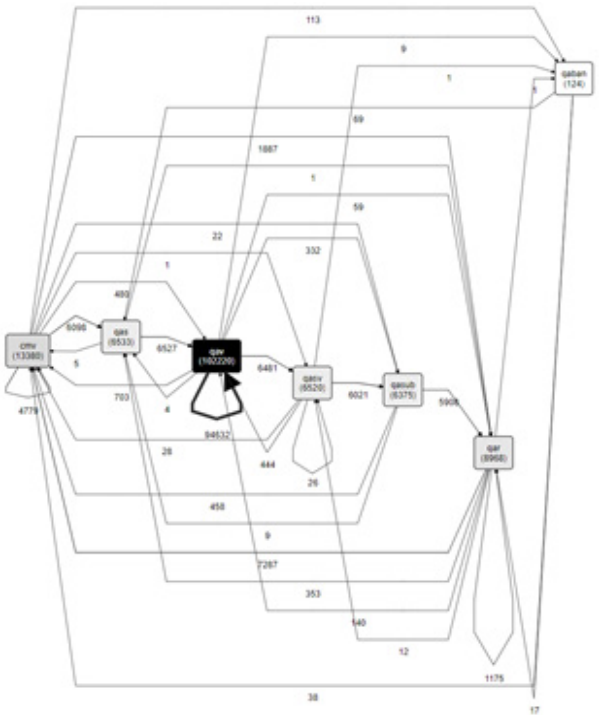


Figure 12. Control flow of scenario 1

By using scenario 1, we can see the visualization of the process from the flowchart. However, we hardly find information from the figure 10 and figure 11. Additionally, from figure 12, the flow of the quiz was observed. It was obvious that the most common pattern is cmv - qas - qav - qasv - qasub - qar. The produced flow was the expected result since it represented the normal attempt of the quiz. For the second scenario, only control flow and dotted charts were shown using event context with a value of Quiz: Quiz 1 – Introduction to Digital World. The calculation of the number of quiz attempts which was conducted during

preprocessing focuses on the starting time of the quiz. The starting time of the quiz was marked with an Event name – labelled *Quiz attempt started* and ended with *Quiz attempt submitted*.

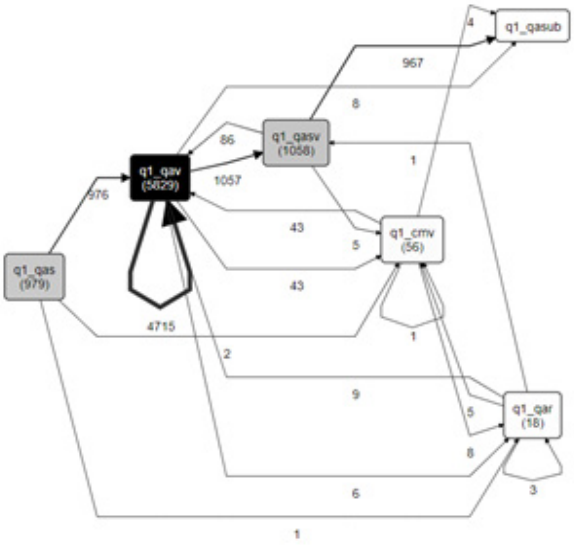


Figure 13. Control flow of scenario 2

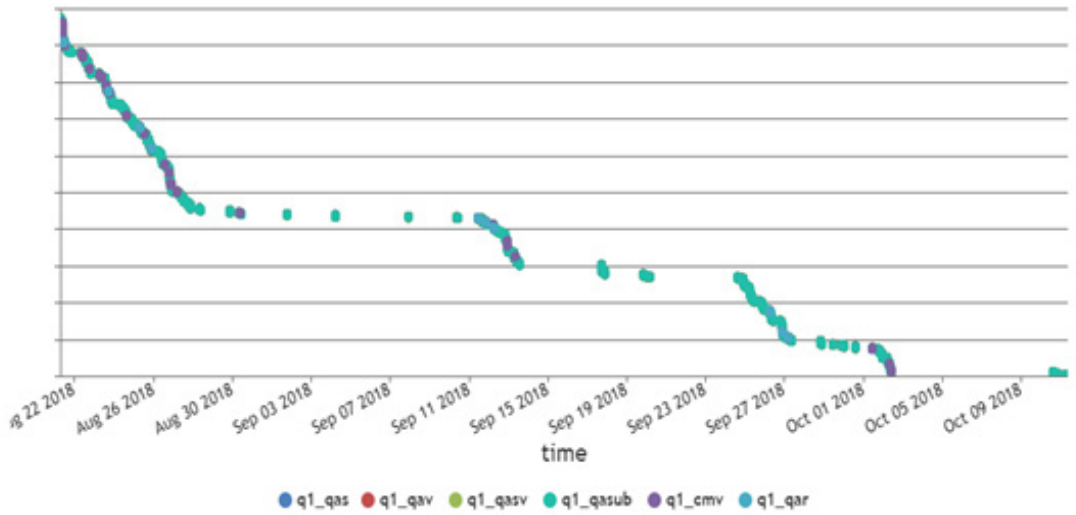


Figure 14. Dotted chart using absolute time in scenario 2

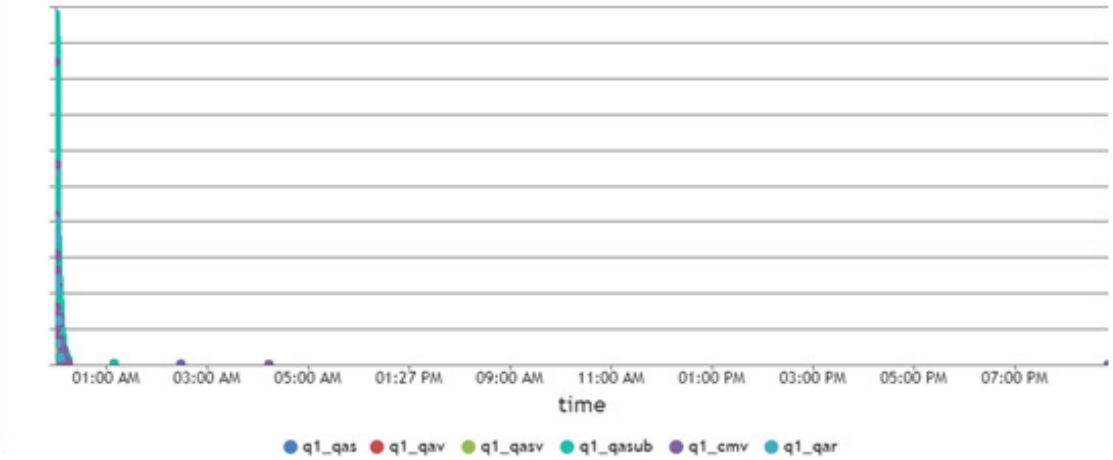


Figure 15. Dotted chart using relative time in scenario 2

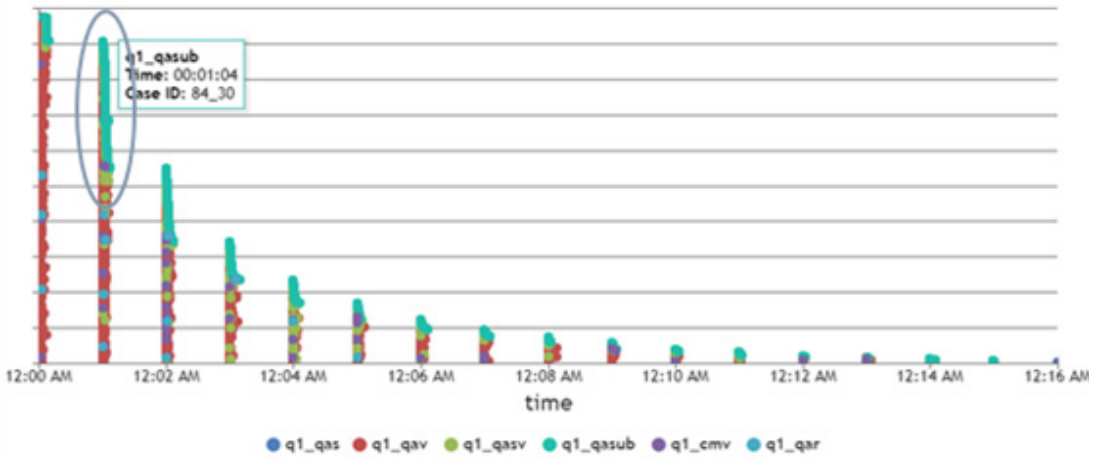


Figure 16. Dotted chart of the results of the anomaly deletion in scenario 2

The control flow using scenario 2 in figure 13, differs from the control flow in scenario 1. The reason was that the scenario 2 focused on the starting point of the quiz – when the Quiz attempt started. It was shown that the control flow obtained from scenario 2 is easier to read since the case id taken focuses on only one quiz. The most common pattern appeared was similar to scenario 1 because it was the expected flow.

By looking at the dotted chart with absolute time in scenario 2 shown in figure 14, it was seen that the highest frequency of quiz 1 is found at the beginning of the lecture in one-week span. Then, the quiz attempt became active again around 11 to 13 September and before the midterm examination (UTS), which was on September 24 to October 2. It was concluded that the participants used the quiz in Moodle to learn and recall the lessons that had been learned for exam preparation.

In figure 15, it is shown that the resulted dotted chart is still quite difficult to read. The reason was that the data anomalies were still occurred in the data – with case id 78_1, 43_1, 60_1 and 26_1. In those case ids, the quiz took more than 1 hour. It should not be possible since the time of the quiz attempt had been set at the beginning. Consequently, data deletion was carried out on the four case ids to earn the more appropriate results. The deletion was done by using the alias naming feature available in preprocessing. The results of the deletion is depicted in Figure 16.

Based on the dotted chart with the relative time in scenario 2 – shown in figure 16, the maximum range of the graph is 100% and each line on the Y-axis had the distance of about 10%. Therefore, it was obvious in the section inside the blue circle, there are around 35% of quiz attempts completed within 1 minute. This was likely due to the quiz resubmission conducted by the participants who had already known the questions after the first attempt and continuously tried to earn better scores in each trial.

4. Conclusion

The conclusion obtained from this study is that the integration of the preprocessing and EDA stages of the process mining was successfully carried out and the system built was able to be used properly. By using this system, teachers and researchers could carry out preprocessing and EDA on Moodle data. For example, based on experiments conducted, it was found that the highest attempt of quiz 1 was found at the beginning and mid-term of the semester. Besides, it was also found that the most time spent in quiz 1 was 1 minute

5. Further Research

For the further research, process mining algorithms will be applied to this system using data that have been processed in this study, so that teachers and researchers can use this system to conduct process mining of data from Moodle. Also, the further research might integrate data retrieval from Moodle into the application by using a web service so that it can reduce the processing time required to export data from Moodle and to upload it back into the system.

References

[1] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*. 2012.

[2] C. Romero and S. Ventura, "Educational Data Mining : A Review of the State of the Art," *IEEE Trans. Syst. Man. Cybern.*, vol. 40, no. 6, pp. 601–618, 2010.

[3] W. M. P. Van Der Aalst, *Process Mining Data Science in Action*, 2nd ed. Heidelberg: Springer, 2016.

[4] K. Grigorova, E. Malysheva, and S. Bobrovskiy, "Information Technology and Nanotechnology," 2017.

[5] A. Bogarín, R. Cerezo, and C. Romero, "A survey on educational process mining," *WIREs Data Min. Knowl. Discov.*, vol. 8, no. February, pp. 1–17, 2018.

[6] D. R. Ferreira, *A Primer on Process Mining*. 2017.

[7] A. M. Momani, "Comparison between two Learning Management Systems : Moodle and Blackboard," *Inf. Syst. Behav. Soc. Methods eJournal*, vol. 2, no. 54, pp. 1–10, 2010.

[8] K. Slaninova, J. Martinovic, P. Drazdilova, and V. Snasel, "From Moodle Log File to the Students Network," 2014, pp. 641–650.

[9] A. Bogarín, C. Romero, R. Cerezo, and M. Sánchez-santillán, "Clustering for improving Educational Process Mining," 2014, pp. 11–15.

[10] A. Bogarín, C. Romero, and R. Cerezo, "Discovering students' navigation paths in Moodle," in *8th International Conference on Educational Data Mining*, 2015, pp. 556–557.

[11] L. Juhanak, J. Zounek, and L. Rohlíkov, "Using process mining to analyze students' quiz-taking behavior patterns in a learning management system," *Comput. Human Behav.*, vol. 92, pp. 496–506, 2019.

[12] K. Willems, "Python Exploratory Data Analysis Tutorial (article) - DataCamp," 2017. [Online]. Available: https://www.datacamp.com/community/tutorials/exploratory-data-analysis-python. [Accessed: 10-Dec-2018].

[13] M. Fani Sani, S. J. van Zelts, and W. M. P. Van Der Aalst, "Repairing Outlier Behaviour in Event Logs," in *International Conference on Business Information Systems*, 2018, vol. 320, pp. 115–131.

[14] N. Tax, N. Sidorova, and W. M. P. Van Der Aalst, "Discovering more precise process models from event logs by filtering out chaotic activities," *J. Intell. Inf. Syst.*, vol. 52, no. 1, pp. 107–139, 2019.

[15] R. P. J. C. Bose, R. S. Mans, and W. M. P. Van Der Aalst, "Wanna Improve Process Mining Results ? It ' s High Time We Consider Data Quality Issues Seriously," in *IEEE Symposium on Computational Intelligence and Data Mining*, 2013, pp. 127–134.