Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

Special Issue

DATA PRE-PROCESSING: A CASE STUDY IN PREDICTING STUDENT'S RETENTION IN MOOC

N. Mohamad, N. B. Ahmad* and S. Sulaiman

Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia

Published online: 05 October 2017

ABSTRACT

Data pre-processing is a crucial phase prior to analytic task and yet rarely been discussed, especially for e-learning data which has multilevel data. Providing a reliable data pre-processing is important to provide quality dataset. Therefore, this study investigates the problems arise in data pre-processing and in this case, for identifying the significant factors to implement prediction task. A MOOC dataset is selected for the data pre-processing task. The process in generating the summary of dataset is explained and the ultimate aim is to produce a dataset with features that are ready for data mining task. The study also proposed a process model and suggestions, which can be applied to support more comprehensible tools for educational domain who is the end user. Subsequently, the data pre-processing become more efficient for predicting student's retention in MOOC.

Keywords: data pre-processing; e-learning; massive open online course; student's retention; prediction.

Author Correspondence, e-mail: bahiah@utm.my

doi: doi: http://dx.doi.org/10.4314/jfas.v9i4s.34

1. INTRODUCTION

Nowadays, massive open online course (MOOC) has become more relevant within learning community. The trend in providing courses is towards supporting lifelong learning and

upgrading professional skills. MOOC is online class that provide free registration for anyone to learn according to their pace and time, with supply of video, quiz and wiki or forum for discussion [1]. Even though there are still people who are not familiar with the term, studies show that students registered with MOOC is increasing [2]. However, despite the great response, low student's retention in MOOC remain a true challenge [3].

Student's retention is the ability of the student to retain from admission until completion of a course [4]. Many studies have been done to investigate and mitigate this issue as student's retention in MOOC is one of the indicator to measure quality of an institutional or the course provided. Therefore, those studies are conducted to increase the completion rates. Also, defray the expenses in providing the course with offering appropriate intervention. In summary, the studies can be concluded; to identify the factors related to student's retention, the improvement of technique or model of prediction or to implement the prediction task and recommend appropriate measurement.

However, as in any other studies, data pre-processing is a crucial phase prior to identify significant factors of student's retention to implement the prediction task and yet rarely been discussed. A lot of time and effort is needed in preparing data with about 70% of time spent and more than 50% of total effort [5-6]. The task become more challenging because of various platforms, features and analysis goals. Moreover, MOOC becomes involve in big data as the number of MOOC students are increasing which a course that generates records every day could reach millions of records in few years [2]. Standard computing environment could not cope to do the analytic task anymore. Therefore, this study presented a process model which employed several model and rules of previous studies. This study discusses the problem arise in data pre-processing for the student's retention prediction task and provide several suggestions during the process.

Data preparation which is part of data pre-processing is a process conducted prior to data mining task, which incorporating several steps from data collection of raw data to format conversion that converts the dataset into readable format [5]. Other term that associates with the data pre-processing task is data wrangling which involves collecting, filling and organizing raw data set and transform the data for later use [6]. Providing a reliable data pre-processing is important to provide a dataset with complete data, reduced noise and minimum error to feeds data mining task the quality data and therefore provides quality result [7].

There are few studies related to data pre-processing, which serve various fields. This paper highlight related studies on educational data including e-learning data. In [8] investigated online reading behaviour. The study provided guide on data pre-processing for identifying frequent patterns of events, which gives inspiration on handling time series data. This study also agrees that implementing data transformation is not a simple task as several events brought to a conceptual meaning or meaningful action thus a single event cannot simply be removed. This issue is addressed which suggest other studies to share experience through research studies. For this study, the process also important for constructing features of time spent which is discuss under next section.

In [9] has provide a very comprehensive guide on preparing data for educational data generally. The study provided steps thoroughly starting from data gathering, aggregation/integration, cleaning, user and session identification, attribute selection, data filtering and data transformation. The study also brought up Moodle as the case study on preparing the data. The process presented, which is for educational data is compatible with this study domain and for student's retention prediction task.

Other study [10] describe rules for reduction process and in [11] give thorough discussion on the nature of edX platform, which is used in this study. Meanwhile, monitoring data quality is also important when talking about processing data. Therefore, in [7, 12] provide comprehensive guide to evaluate data quality.

From those studies, several models can be adapted which are Romero method for the whole preparation process generally. However, the arrangement of the process is considered to fit this study purpose. Also, this study includes Kotslantis [10] rules and Parashar [11] rules for instance or attribute selection task and added element of feature construction. Next section discusses on each process using dataset collected, followed by discussion (discuss on evaluation and issues) and conclusion**.**

2. METHODOLOGY

Every fields may need different type of data for the purpose of data mining task. This also applied to educational data, which may also require the same type of data even though they serve different educational problems. Also, differed with other fields, e-learning data has "multiple levels of meaningful hierarchy" [8]. Table 1 shows features that needed to be prepared for predicting student's performance, which is summarized from various studies. The features are derived from actions of MOOC activities which also common for e-learning.

The activities include participate in discussion forum, doing quiz or test, login account, browse and view content, checking progress and assignment.

The features shown in Table 1 concluded from various studies that proven to be significant or contribute to student's retention or drop out in e-learning. The features can be categorized into several types which include total number of actions (frequency) and the time spent on an action (duration). The data can be concluded to require time series data. In [9] case study on Moodle platform, the data cannot be acquired directly from the edX platform but must go through the feature construction process which is discussed under activities implementation section. Therefore, the next section will discuss on the preparation and provision for those data with several features is discussed as an example.

Table 1. Type of data needed for student's retention prediction task

Activities	Actions	Sources
Forum	•View post •Add post (original and follow-up) •Time spent add post (original and follow-up) •Post length	[13-20]
Quiz	•Number of attempts •Succeed the first quiz	[13-14, 18, 21]
Login	•Login frequency per day •Time spent online	[12, 21]
Content	•Number of clicks for main page and materials •Time spent on view content pages	[15, 17-20, 22]
Progress	•Number of time checking progress	[17, 19, 22]
Assignment	•Time spent to complete •Total score •Times read •Duration to datelines	[19, 22]

Fig. 1 shows the process model applied for the data pre-processing. The process model includes data collection, data split, user identification, activities implementation and transformation. In this model, instead of focus on the arrangement of data process, the model focus on the activities or the goal that need to be achieved. For example, this study focuses more on the construction of statistic or the construction of time series data which consist of the data process like filter and data reduction.

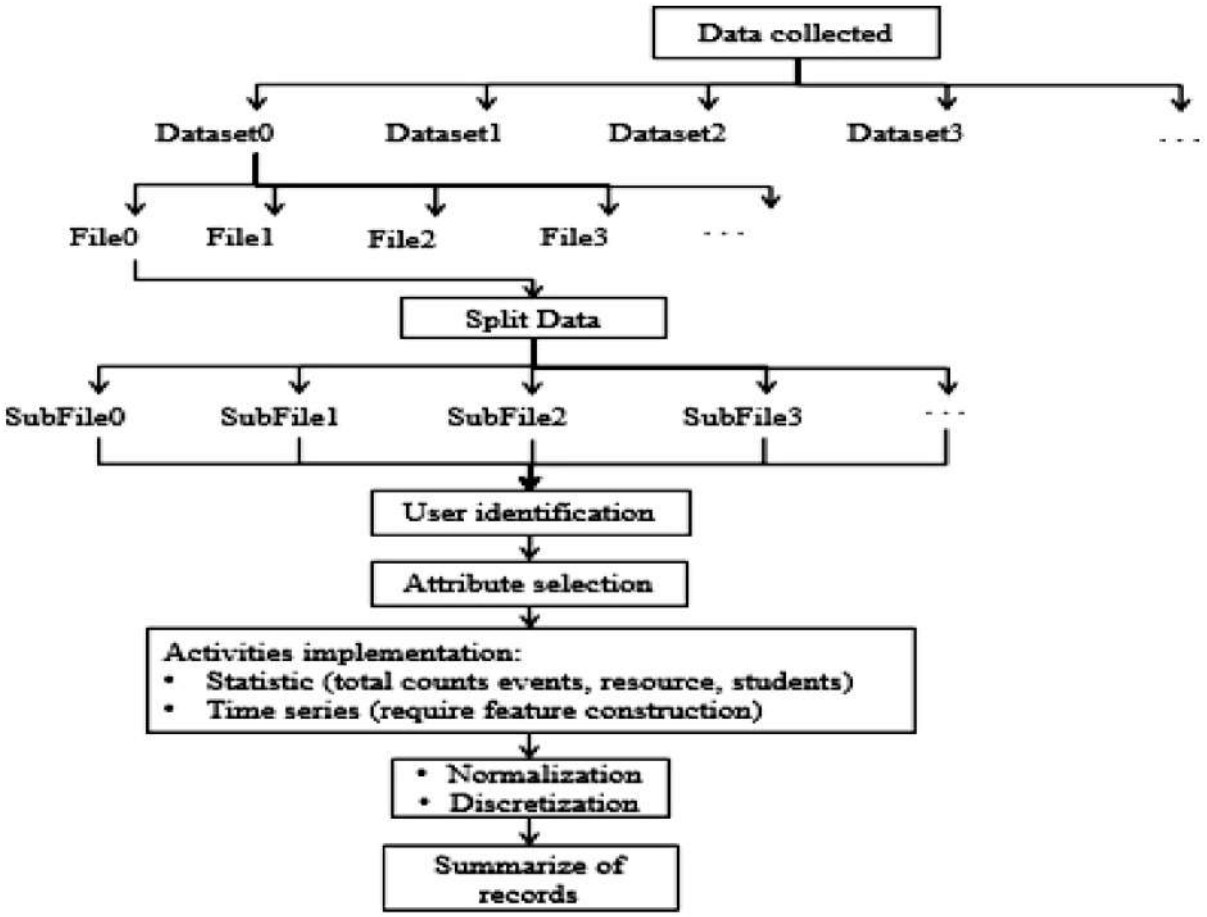

Fig.1. Process of data pre-processing

The focus on activities implementation is because a process might be required by several activities like removal of unqualified records is done in both constructions of statistic and time series data. In [7, 12] explain that the arrangement of data pre-processing has become an open issue. Moreover, for e-learning data, the process need to be defined by the domain expert who ask the question to avoid any lose context meaning [8]. The activities are focus on after the data is split into several files for faster processing. The following section will discuss the whole process in more detailed starting from data collection.

2.1. Data Collection

A course with the latest 14753 students with 2.4 million records is selected for pre-process. The other larger MOOC dataset could reach more than 40 million records. However, for this study, the dataset which is larger than usual e-learning class size (around 20 to hundreds of students [23]) is selected as an example. This study acquired approval from CAROL, Stanford University to obtain and process the data. Therefore, this study has no experience to discuss the issue raise from extracting the data from the institution database. However, several issues have been explained in the data documentation. The course is conducted on OpenEdX which

is one of the most popular MOOC platform and OpenEdX is using MongoDB for database program.

From the data collected, for each course, there are several types of files which included but not limited to eventXtract (containing student's events or actions), demographic (brief description on student's background), CourseInfo (a row of information about the course like start and end date), VideoInteraction (events of video with more detailed) and several others. Explanation on the data are usually documented by the provider for references. Also, there is MOOCdb project [24] that provide comprehensive guide for data model and a collaborative research platform. Starting from this point, identifying the right tools for pre-process is important. The options include whether to employ combination of tools, using a complete packaged software or write code on available platform.

2.2. Data Split

After acquiring data and understanding the data structure, the process which distinguish the standard analytic project and big data project is data split process which applying the concept of map and reduce. In map and reduce concept, the master distributes the idle task to the slave [25]. A large file which is infeasible for standard processing tools is distribute to several files. The process reduces the burden of computation memory. This process can be achieved by using a packaged software or platform that support the map and reduce process along with the whole process of data pre-processing. A few examples of this tool or platform are Apache Hadoop, Apache Spark and the latest Apache Flink [12].

For other option, a specific tool can be used to split a large file into partition. This method is to have better speed at data loading (especially for limited memory and limited network performance) [26]. Even though using the specific tool requires computer storage, the task is more convenient for small and personal educational research project as it requires less technical skills and more privacy. Example of the specific tool that split a file are csv splitter, csv chunker and csv utility. For an example in this case, a file of event_type can be split for each million records and generate three files. After data is split, the process on activities can be executed.

2.3. User Identification

User identification is required to identify individual who involved in the study. Research need to know which students need to be included in the later analytic task, according to research purpose. Unique id is created because for this case, in the event_type file, the respondId or student id information are the unique id available to identify a student. User also can be

identified through anon_screen_name, which is anonymous screen name chosen by student. However, this study could not ensure if the name is unique because there are several logs with same name but belong to multiple countries. The name might belong to different students or the same students who travel a lot.

2.4. Attribute Selection

Attribute selection is "to identify and remove as much irrelevant and redundant information as possible" [9]. This subsection using the term attribute selection to distinguish with feature selection that will be used at the end of this section. Attribute selection process select only relevant attributes for the study such as excluding all other video event (video speed, video new speed, video current time and video old time) which is not the focus of this study. The process is important for e-learning including MOOC, which has many attributes comprised in a file [12]. This process makes the later work easier with less memory consumption and researcher can focus more on objective of the task.

2.5. Activities Implementation

After user identification, it is easier for choosing records to extract and construct features which is implemented under activities. The activities objective is to produce data that is needed according to research studies. For this study, the objective of the activities is to produce statistic and time series data.

2.5.1. Statistic Construction

Statistic construction requires filtering process and the goal of this task is to know how many students involved, how many activities they need to complete. Sometimes, studies that investigate student's retention use grades while others use completion rate as indicator or dependent variable for the student's success. For example, as shown in Fig. 2, after split data for each file, data with duplicate rows is filtered for unique records and result of all files is integrated in one file. After the integration, the file is filter again and generate summarization of the records. Meanwhile, Fig. 3 shows example of type of resources exist in a file and being filtered to produce the unique records only.

Table 2 shows example of summary of records that might be relevant to understand the dataset. The information includes number of type of event available, the latest number of student participates and number of resources provided. In this study, information on resources give insight on how many topics the students need to learn and how many quizzes they need to answer to complete the course. This task also can be done to check the information provided by the course information file if one is given.

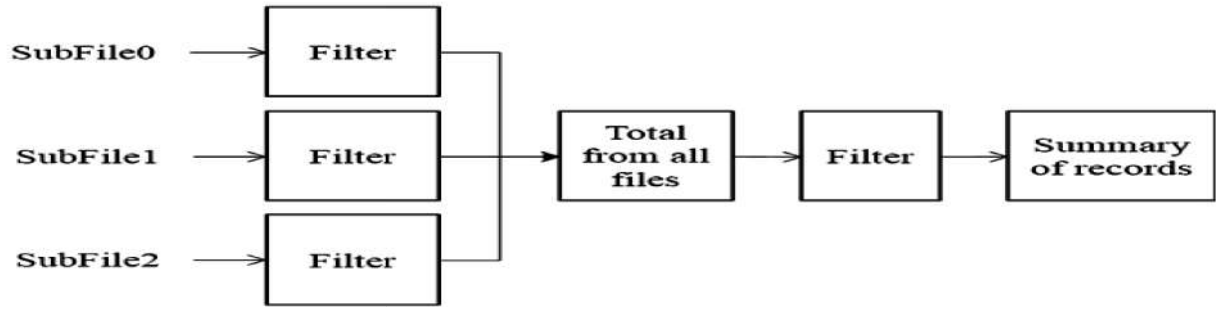

Fig.2. Process of constructing statistic

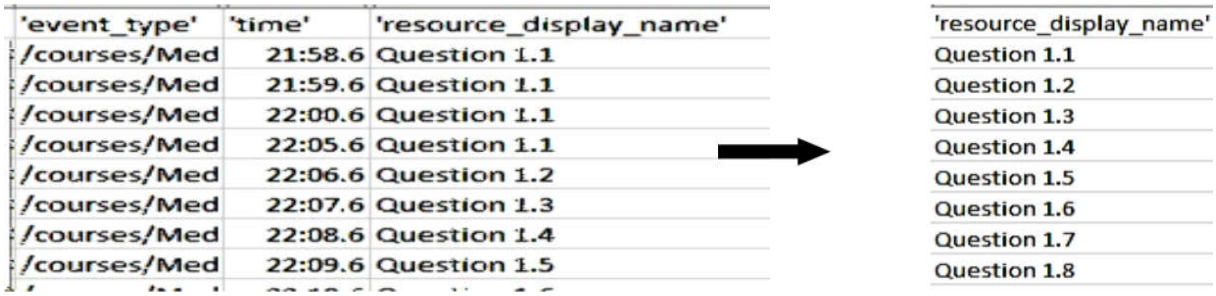

Fig.3. Filtering process for constructing statistic

Table 2. Statistic of records

Categories	Total
Events	1235
Students	13734
Resources	273 (include 8 topics, 52 videos and 90 questions from quizzes/tests)

2.5.2. Time Series Data Construction

Time series data construction requires feature construction. Fig. 4 shows the process for the construction. First, required attributes are selected from each file like id, event, date and time and the construction of duration. Then, illegal values are remove according to Kotslantis rules [10] like error and value with more than permissible range. For example, a video event that has duration for too long which is 11 hours can be removed because there are no videos with that length in the list. This task can be initialized with the features of frequency first (such as number of post, number of view and number of time checking progress).

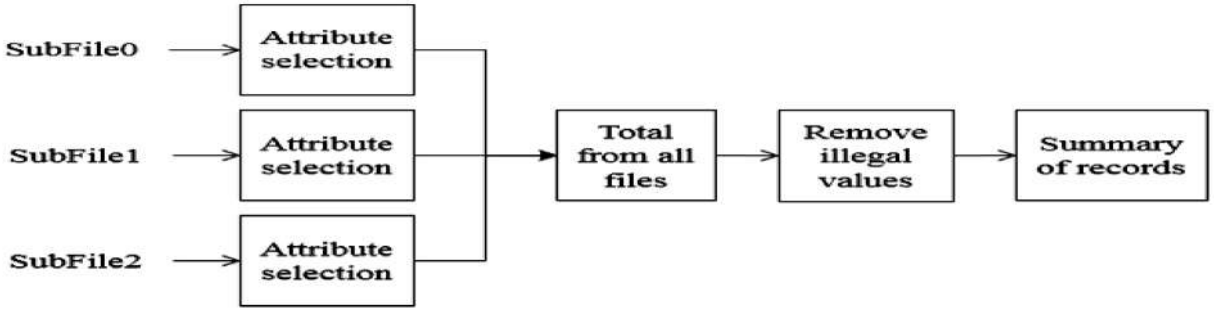

Fig.4. Process of constructing time series data

Then, the task can proceed with feature of time spent such as time spent create forum post, time spent view content and duration between assignment submission and due date. For the purpose of constructing feature of seconds on view content page (view topic and view video), duration of video is calculated starting by the video load and end. For this calculation, in Parashar rules [11] can be applied which removing in between video event like play_video which occur automatically for every 2 minutes. Also, other several events can be removed like seek_video and seek_change video. However, pause_video with its next event need to be retained for the reduction of the duration because only time spend on viewing video need to be considered. These made the records cannot simply be removed at the first phase. Table 3 is example of video event and the steps involved is depicted in Table 4 and is discussed next.

Table 3. Example of video event

Event_Type	Date Time
load_video	3/4/2016 6:10:29 PM
play_video	3/4/2016 6:10:31 PM
seek_video	3/4/2016 6:11:18 PM
pause_video	3/4/2016 6:13:11 PM
play_video	3/4/2016 6:18:29 PM
seek_video	3/4/2016 6:18:12 PM
stop_video	3/4/2016 6:18:29 PM

Table 4. Duration for feature of seconds on view content

Event_Type	Date Time
load_video	T1
play_video	T2
seek_video	T3
pause_video	T4
play_video	T5
seek_video	T5
stop_video	T6

Based on [8], the sequence can be described as below:

$\alpha = < T1, T2, T3, T4, T5, T6, T7 >$

$\beta = < T1, T4, T5, T7 >$

where β is complete sequence for viewing a video. After removal, the new sequence is α which the information is used for calculation of dv (duration of video). Meanwhile, the calculation of duration can be concluded as below:

$$dv = (T7-T1) - (T5-T4) \quad (1)$$

where T7 is the last time of event in the sequence minus T1, which is the starting time of event. While for calculating the pause_video duration, T5 minus T4. T5 is the next event of T4, the pause_event. This might be different for other sequence of video event as they have different length of sequence. Thus, identifying sequence and calculate duration for each action need to be done cautiously as the task is challenging.

Moreover, the process of seconds on view features, can only be performed after feature of login frequencies has been constructed. This is because the event like page_close is required to identify number of time the student login (in case such information is unavailable). Feature of login frequency requires the page_close event to count the frequency.

All the features constructed are fill in the new table as shown in Table 5. This table are constructed for every week as for prediction, several studies [3] compare the performance of students from several weeks of the course.

Table 5. Table for time series data

Week1

ID	V1	V2	V3	V4	V5	V6	V7	V8	…
S1									
S2									
S3									
. . .									

2.6. Data Transformation

After the dataset is prepared, normalization is needed to adjust the data to follow normal distribution as the variables are disparate with various measurement. There are various techniques for normalization like binarizer, normalizer and standardscaler [12]. For e-learning variables, common technique like min-max can be used [9]. The technique converts all the data into values with range for example, 0 to 1.0. Normalization is also recommended if neural network will be used for data mining technique.

Meanwhile, discretization is required when the researcher need to scale continuous data into class that is useful and easier for understanding the result. For example, continuous values of number of post is scale into zero (0), low (1-5), medium (6-10) and high (> 10) which can be keyed in as 0, 1, 2 and 3. Several techniques can be used for discretization which include

bucketizer, discrete cosine transforms and elementwise product [11]. Other techniques include using supervised and unsupervised algorithm.

Apart from that, feature selection is done to identify and selecting the most significant features before implementing the data mining task such as classification. This task is important which many studies that investigate student's retention, focus on the feature selection to find which features are the most significant for the prediction of retention or drop out. There are many techniques for feature selection like vectorslicer, RFormula and Chi-squared selector [12]. Decision tree technique can also be used like [27] study, which use the technique in predicting whether student will pass or fail.

After completing all the process, the format conversion can be done. Format conversion involves converting the dataset into readable format (such as .txt, .xls, .csv, .arff or .xml) to be used by analytic tools or platform. Then, the dataset is ready for analytic and visualisation.

All the process discussed previously can be support by tools which are developed by dedicated research group and commercial company. Examples of popular tools are OpenRefine by Google, Trifacta, R, SPSS, SAS and Rapidminer. Even the spreadsheet tool like excel is equips with PowerQuery tools that can process millions of data. The tools save analyzer time and effort especially for big data which many of the tools or platforms conduct the process on the cloud so user don not have to invest much on storage. Next section summarizes all the issues found during the data pre-processing.

3. RESULTS AND DISCUSSION

Investigation on student's retention in MOOC can help those thousand students completing a course successfully. Teacher need to get insight on what they do, factor that influence them learning and how they learn in online learning to understand them. To do that, analytic is needed. However, it is challenging for even the data pre-processing task. Preparing data before implementing prediction task is long and tedious works. Several issues are found during the task which is discuss next.

3.1. Issues in Selecting Data and Removing Duplication

When selecting data, there are some students who might not really want to learn the whole course which means they just want to learn selected topics. Even more, there are some students who want to see if the course is suitable for them. Research might want to exclude this group of students. Unless there is data from survey of student's intention, it is not easy to observe, select and filter and interpret this type of students from a large dataset.

In term of removing duplication, in e-learning, duplicate records are rarely an issue. However, there are multiple records which generate by the same student but refer to different action. This records cannot simply be deleted because depending on the analysis goal, the duplication need to be counted. For example; the records could mean total number of hits for each activity or sum up of duration of activities.

3.2. Issues in Activities Implementation

E-learning including MOOC has many variables from various dimensions. For example, a dataset may contain many students with various variables for different period with grade achievement. Not to mention those variables are filled with values which are inconsistence and messy. Therefore, delay in the cleaning process for feature extraction and construction which need to consider an action context meaning.

There are various activities in e-learning, which researcher need to define carefully. Some actions might contribute to the same features. For example, like viewing subtopic means receiving the same content as watching video. Some courses, provide the same content of a topic as the video provided to ease students. This means if a student not watching certain video, he/she might prefer and have read it through the page for the topic. Therefore, this study considers view video and seconds on video as view content and second on view content. Time taken to complete a course are different for each student and unpredictable. Some students drag time to complete a course for too long and some not even return even though they did not click unenrolled button. The unclear situation of student returns with the long duration of course, make labelling class task challenging.

If time spent on activities need to be analyzed, calculating the duration is challenging as there is no specific guide on dictating the start and end of activities. In [8] state that there are no universal rules to follow. For example, in event activities file, it is unknown if a student already watched a video because in some case the video may set to auto loading. Moreover, this not include the issue arise if the research seeks the factor of student's particular behaviour on video. Like a situation where student jump to certain part of the video to investigate which part of video has the students viewed.

In summary, the most challenging part during data pre-processing is the feature extraction and construction process under construction of time series data during activities implementation. As mentioned before and in other studies [8, 12], the researcher need to know how to define each data process in data pre-processing. Therefore, the data process model need to be established thoughtfully.

Before conclude, this study would like to discuss on the evaluation of data. Evaluation of data is important to ensure the dataset achieve good level of quality and ready to be analyzed or for reuse. This task is recommended after the collection of data. The dataset can be evaluated based on [7] model (DOP8_Qpp) using UCUA which is utility, compliance, usability and acceptability. Meanwhile, in [28] also mentioned those dimensions under intrinsic category while under context category, the dimension include reputation, accessibility, believability and value-added quantity.

For example, under utility dimension, after collecting data, it is important to check all required data are there with variables needed to answer research question. There is possibility of discrepancy like unsynchronized data among data files. For example, a student who play a video is existed in event file but cannot be found in self-pace video interaction file. This situation might occur for system or database which still under development or maintenance.

Other dimension is compliance which consist of completeness, consistency and accuracy. For compliance, researcher need to identify which datasets have missing value. Apart from that, researcher must make sure the data are accurate or make sense. Records with unusual values like date time that outside of course established can be removed. Also, the structure of data is in correct format.

Another example of dimension is timeliness which require researcher to check, whether the data are recent enough or up-to-date. Data need to be up-to-date because most of the analytic purpose is to answer question proposed. Thus, feeding the analytic with recent data provide more accurate and up-to-date result. Finally, the study is concluded in the next section.

4. CONCLUSION

This study presented the process on data pre-processing for a MOOC dataset. Data pre-processing is crucial phase before implementing analytic task. The study focus on activities implementation and applied methods from previous studies, which incorporate process from data collection to data transformation and some rules is discussed in detailed for feature construction. The solution for issues like selecting the right data, design the right steps for reduction and features construction, need to be discussed in future study. Lastly, model for data evaluation is applied to make sure the data is in good quality and ready to be used for educational research like learning analytic and educational data mining.

5. ACKNOWLEDGEMENTS

The authors would like to express our deepest gratitude to Research Management Centre (RMC), Universiti Teknologi Malaysia (UTM) for the support in R&D (FRGS, 4F496) and Soft Computing Research Group (SCRG) for the inspiration in making this study success. Also, we would like to thank Center for Advanced Research through Online Learning (CAROL), Stanford University for the assistance and opportunity to access and explore the data.

6. REFERENCES

[1] Hoy M B. MOOCs 101: An introduction to massive open online courses. Medical Reference Services Quarterly, 2014, 33(1):85-91

[2] Eichhorn S, Matkin G W. Massive open online courses, big data, and education research. New Directions for Institutional Research, 2016, 2015(167):27-40

[3] Khalil H, Ebner M. MOOCs completion rates and possible methods to improve retention-A literature review. In World Conference on Educational Multimedia, Hypermedia and Telecommunications, 2014, pp. 1305-1313

[4] Fenty D, Messemer J, Rogers E. Adult entrances and exits: What does retention literature inform us about urban adult higher educational participants and student success? In Adult Education Research Conference, 2016, pp. 278-279

[5] Ramírez-Gallego S, Krawczyk B, García S, Woźniak M, Herrera F. A survey on data preprocessing for data stream mining: Current status and future directions. Neurocomputing, 2017, 239:39-57

[6] Terrizzano I G, Schwarz P M, Roth M, Colino J E. Data wrangling: The challenging yourney from the wild to the lake. In 7th Biennial Conference on Innovative Data Systems Research, 2015, pp. 1-9

[7] Mandran N, Dupuis L, Luengo V. DOP8_Qpp: Model to pre-process educational data. 2016, https://hal.archives-ouvertes.fr/hal-01321345/document

[8] Zhou M. Data pre-processing of student e-learning logs. In Information Science and Applications, 2016, pp. 1007-1012

[9] Romero C, Romero J R, Ventura S. A survey on pre-processing educational data. In A. Peña-Ayala (Ed.), Educational data mining. Cham: Springer, 2014, pp. 29-64

[10] Kotsiantis S B, Kanellopoulos D, Pintelas P E. Data preprocessing for supervised leaning. International Journal of Computer Science, 2006, 1(2):111-117

[11] Parashar R D. Student performance analytics for blended MOOCs on IITBombayX. Master thesis, Maharashtra: Indian Institute of Technology Bombay, 2016

[12] García S, Ramírez-Gallego S, Luengo J, Benítez J M, Herrera F. Big data preprocessing: Methods and prospects. Big Data Analytics, 2016, 1(1):1-22

[13] Khalil M, Ebner M. Learning analytics in MOOCs: Can data improve students retention and learning? In World Conference on Educational Media and Technology, 2016, pp. 581-588

[14] Gitin E, Niemi D, Saxberg B. Online student behaviors and retention. In World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, 2015, pp. 785-792

[15] Wolff A, Zdrahal Z, Herrmannova D, Kuzilek J, Hlosta M. Developing predictive models for early detection of at-risk students on distance learning modules. Machine Learning and Learning Analytics Workshop at the 4th International Conference on Learning Analytics and Knowledge, 2014, pp. 1-4

[16] Yang D, Sinha T, Adamson D, Rosé C P. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In NIPS Data-Driven Education Workshop, 2013, pp. 1-8

[17] Balakrishnan G, Coetzee D. Predicting student retention in massive open online courses using hidden Markov models. Technical Report No. UCB/EECS-2013-109, Berkeley: University of California, 2013

[18] Wolff A, Zdrahal Z. Improving retention by identifying and supporting" at-risk" students. EDUCAUSE Review Online. 2012, http://er.educause.edu/articles/2012/7/improving-retention-by-identifying-and-supporting-atrisk-students

[19] Lykourentzou I, Giannoukos I, Nikolopoulos V, Mpardis G, Loumos V. Dropout prediction in e-learning courses through the combination of machine learning techniques. Computers and Education, 2009, 53(3):950-965

[20] Morris L V, Finnegan C, Wu S S. Tracking student behavior, persistence, and achievement in online courses. The Internet and Higher Education, 2005, 8(3):221-231

[21] Herrmannova D, Hlosta M, Kuzilek J, Zdrahal Z. Evaluating Weekly Predictions of At-Risk Students at The Open University: Results and Issues. In Annual Conference Expanding Learning Scenarios: Opening Out the Educational Landscape, 2015, pp. 200-208

[22] McCuaig J, Baldwin J. Identifying successful learners from interaction behaviour. In 5th International Conference on Educational Data Mining, 2012, pp. 160-163

[23] Chen B, Zydney J, Patton K. Creating a community of inquiry in large-enrollment online courses: An exploratory study on the effect of protocols within online discussions. Online Learning, 2017, 21(1):165-188

[24] Veeramachaneni K, Halawa S, Dernoncourt F, O'Reilly U M, Taylor C, Do C. MOOCdb: Developing standards and systems to support MOOC data science. 2014, https://arxiv.org/pdf/1406.2015.pdf

[25] Zha L, Zhang J, Liu W, Lin J. An uncoupled data process and transfer model for MapReduce. In A. Hameurlain, J. Küng, R. Wagner, L. Bellatreche, & M. Mohania (Eds.), Transactions on large-scale data-and knowledge-centered systems XVII. Berlin: Springer, 2015 pp. 24-44

[26] Sahoo P R, Phalak C. High speed file splitter: A solution to address data split need for parallel processing. In Annual IEEE India Conference, 2011, pp. 1-6

[27] Wolff A, Zdrahal Z, Nikolov A, Pantucek M. Improving retention: Predicting at-risk students by analysing clicking behaviour in a virtual learning environment. In ACM 3rd International Conference on Learning Analytics and Knowledge, 2013, pp. 145-149

[28] Taleb I, Dssouli R, Serhani M A. Big data pre-processing: a quality framework. In IEEE International Congress on Big Data, 2015, pp. 191-198

How to cite this article:
Mohamad N, Ahmad NB, Sulaiman S.Data pre-processing: a case study in predicting student's retention in MOOC. J. Fundam. Appl. Sci., 2017, 9*(4S), 598-613.*