Uncovering Learning Patterns in a MOOC through Conformance Alignments

Patrick Mukala, Joos Buijs, and Wil van der Aalst

Department of Mathematics and Computer Science
Eindhoven University of Technology, Eindhoven, The Netherlands
{m.p.mukala,j.c.a.m.buijs,w.m.p.v.d.aalst}@tue.nl

Abstract. Web-based learning is now offered in multiple forms. One of these is the phenomenon of Massive Open Online Courses (MOOCs). Several approaches in Learning Analytics (LA) attempt to analyze and explain students learning patterns in MOOCs. In addition to traditional data mining techniques, online surveys constitute another way used in LA for analyzing students' learning habits in MOOCs. However, such an approach can be error-prone with data collection. Therefore, we adopt the use of process mining techniques. Process mining techniques provide more robust ways of extracting, analyzing and visualizing students' activities trail. In this paper, we make use of alignment-based conformance checking to extract and analyse students' learning patterns in MOOCs. The aim is to provide a guideline and demonstrate how process mining can provide critical insights in tems of students' learning and quiz submissions behavior, their resulting performance and the correlation therein.

Keywords: Learning Analytics, Mooc, Coursera, Educational Data Mining, Process Mining,Online Learning,Conformance Checking

1 Introduction

Massive Open Online Courses (MOOCs) generate significant data about students. Using this information, several approaches in Learning Analytics (LA) attempt to analyze and explain students learning patterns in MOOCs [6,5,7]. While using traditional data mining techniques, these appoaches also employ online survey and questionaires for data collection [7]. Such approaches are error-prone and can lead to incorrect findings. Therefore, we believe that it is more effective to directly analyze data from students' activities as they interact with videos, submit quizzes and participate in forums. This can be realized by means of process mining techniques [1,11].

Process mining techniques [1] provide more robust ways of extracting, analyzing and visualizing students' activities trail. In [8], we demonstrated some exploratory results and visualized a number of behavioral charateristics using process discovery, dotted chart and conformance checking[4,1].

However, it is possible to go further based on such analysis and to discuss learning patterns exhibited by these students. In particular, with conformance checking [2] we can produce a diagnosis of students' behavior and quantify different behavioral traits among them.

In this paper, we make use of alignments and diagnostics details from alignment-based conformance checking[4] to detail students' learning patterns in a MOOC. The primary objective is to provide a guideline and demonstrate how process mining can provide critical insights in tems of students' learning and quiz submissions behavior, their resulting performance and the correlation therein. By exploring alignments, we can determine how students watch individual videos, which videos are skipped, what the interval for watching successive videos is etc. By getting such detailed information, it becomes easier and more helpful to focus on specific elements, i.e. a video that has been watched regularly, a submitted quiz or a skipped video, and devise necessary interventions in order to improve both the students' learning experience and contents delivery. We therefore make use of a case study to demonstrate how this works.

The remainder of this paper is structured as follows. We discuss all the preliminaries to this study by introducing our Coursera case study and succintly describing the reference model for the analysed data in Section 2. We also give a brief summary of how we generate the event log, the normative model as well as a description of conformance checking. In Section 3, we give and describe a few rules used to specify both the video watch status and viewing habit. We discuss some results in Section 4. Section 5 concludes this paper and discusses possible future directions.

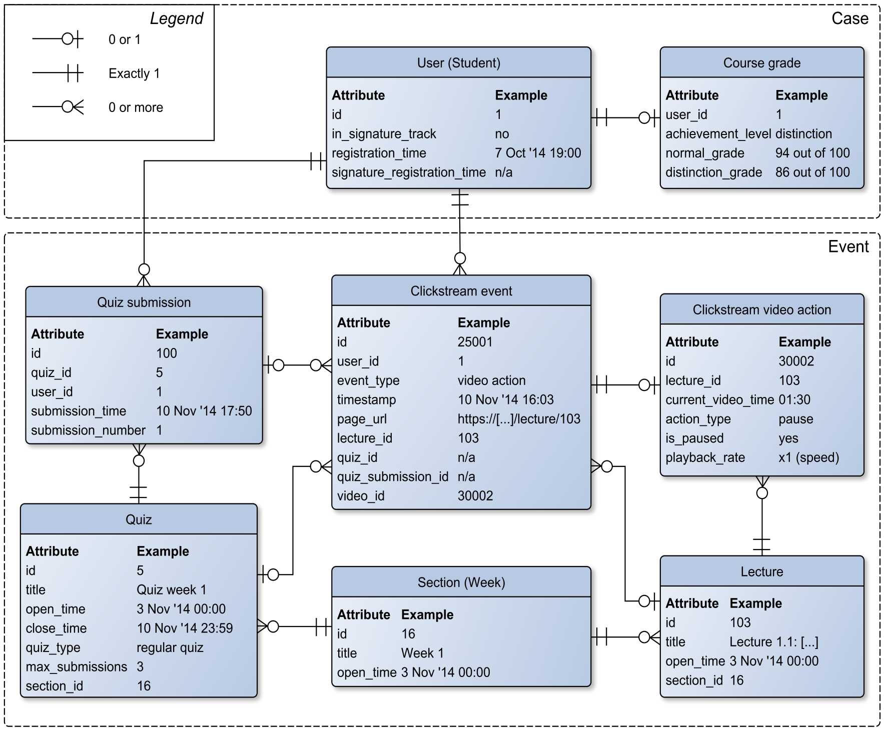
2 Preliminaries

2.1 Case Study

We make use of the same case study used in [8]. The datasets we analyse were obtained from Coursea for the MOOC "Process Mining: Data Science in action" which ran from November 11, 2014 to January 8, 2015. In the end 43,218 students registered, of which 20,868 watched at least one lecture, 5,798 students submitted at least one exercise and 1,688 certificates were issued. These datasets are centered around the *students* participating in a MOOC, and the stream of *click events* they generated on the course webpages. The structure of this dataset is shown in Figure 1, and we will discuss it in more detail below.

Clickstream During a course, students visit the course website to, amongst other things, watch lecture videos and make quizzes. As students click through the website to look up these videos and quizzes, they leave a trail of *click events*, collectively called a *clickstream.* Each such event could be associated with, for example, a particular lecture, or a particular quiz submission. In addition to the pages visited by a student (recorded as a *pageview action*), we also know how the students interacted with the lecture videos (recorded as a *video action*).

Student For each student, we have information about when the student registered for the course, and their end course grade. For the registration, we know the exact *time the student registered*, and if they participated in the special



User (Student)

Attribute	Example
id	1
in_signature_track	no
registration_time	7 Oct '14 19:00
signature_registration_time	n/a

Course grade

Attribute	Example
user_id	1
achievement_level	distinction
normal_grade	94 out of 100
distinction_grade	86 out of 100

Quiz submission

Attribute	Example
id	100
quiz_id	5
user_id	1
submission_time	10 Nov '14 17:50
submission_number	1

Clickstream event

Attribute	Example
id	25001
user_id	1
event_type	video action
timestamp	10 Nov '14 16:03
page_url	https://[...]/lecture/103
lecture_id	103
quiz_id	n/a
quiz_submission_id	n/a
video_id	30002

Clickstream video action

Attribute	Example
id	30002
lecture_id	103
current_video_time	01:30
action_type	pause
is_paused	yes
playback_rate	x1 (speed)

Quiz

Attribute	Example
id	5
title	Quiz week 1
open_time	3 Nov '14 00:00
close_time	10 Nov '14 23:59
quiz_type	regular quiz
max_submissions	3
section_id	16

Section (Week)

Attribute	Example
id	16
title	Week 1
open_time	3 Nov '14 00:00

Lecture

Attribute	Example
id	103
title	Lecture 1.1: [...]
open_time	3 Nov '14 00:00
section_id	16

Fig. 1: The structure and type of information used in our analysis, described in an Entity-Relationship Model.

(paid) *signature track*, in order to obtain a verified certificate. The course grade consists of two parts: the *normal grade* and the *distinction grade*. In addition, the student is assigned an *achievement level* based on the obtained grades. If the student did not complete the course exams, the achievement level is *absent*. If the student did complete the exams, but his normal grade was not sufficient, the student *failed* the course. On the other hand, if the student did have a sufficient normal grade, but insufficient distinction grade, they get the achievement level *normal*. Finally, if the student both has a sufficient normal and distinction grade, they achieved the level *distinction*.

Course structure Lastly, in a Coursera MOOC, lectures and quizzes are grouped into *sections*, (typically *weeks*). Each section is visible to the students at a predetermined time (the *open time*), in order to give structure to the course. Within a section, lectures and quizzes may have their own open time, to further guide students to follow a particular study rhythm. Finally, quizzes an also have deadlines (the *close time*), and quizzes can be attempted multiple times by the student, up to a certain *submission maximum*.

2.2 Building the Event Log

In this case study we are interested in analysing students' behaviors based on the trails of *click events* they generated. Before we can use *process mining* to analyze this behavior, we first need to map the MOOC data to an *event log*. There are two things we must specify for this mapping: what constitutes an *event*, and what makes a *case* (i.e., a sequence of events).

As we are focussing on the behavior of students, we will consider each student as an individual case. The clickstream a student generated will be the basis for the events in this trace. This separation between case and event is also displayed in Figure 1. For this analysis, we will primarily focus on events with the type *pageview action*.

As an example, consider the event log is shown in Table 1. For each case, we store the data available about the student, including their course grade data. For each clickstream event, we create an event belonging to the corresponding case (based on the student user_id). In this example, we will only consider lecture pageview actions. That is, we filtered the MOOC data to get a view of the lecture watching behavior of students. For each clickstream event, we store the click event data, including the referenced lecture as event name.

Based on different data attributes we can determine several students groups. First of all, we can group students that failed (F) the course or successfully (S) obtained a certificate, which can be split into a normal (N) certificate or certificate with distinction (D). The second attribute on which we can split is whether a student enrolled in the signature track (T) or not (F). Thirdly we can consider for which weeks events were recorded, e.g. for week one only (1), weeks one and two (2), weeks one, two and three (3), or all.

Table 1: Example of a generated event log

Cases

id	user_id	in_signature_track	registration_time	achievement_level	...
1	1	no	7 Oct '14 19:00	distinction	
2	2	no	9 Oct '14 01:05	failed	
⋮					

Events

id	case_id	clickstream_id	event_name	timestamp	...
1	1	25000	Lecture 1.1: [...]	10 Nov '14 16:01	
2	1	25002	Lecture 1.2: [...]	10 Nov '14 16:42	
3	2	25003	Lecture 1.1: [...]	11 Nov '14 02:05	
4	2	25004	Lecture 1.2: [...]	11 Nov '14 02:15	
⋮					

2.3 Conformance Checking

Given a normative model such as the one in Figure 3 and a log in Table 1, we can perform conformance checking and verify a score of measurements. Several algorithms address such measurements with their relative limitations. There are several mesurements that can be verified through conformance. Rozinat et al. [10] enumerate four dimensions to consider for determining the adequacy of a model in describing a log: fitness, precision, generalization and simplicity. Selected studies addressing these dimensions include the work in[3,4,9].

Within the bounds of this paper, we relay on the alignments generated as a result of measuring both fitness and precision following alignment-based conformance as summarised in Figure 2.

We perform conformance checking to quantify the watching behavior for these groups over the duration of the course. Making an assumption that all students follow the course in sequence, we designed a model to represent this hypothesis. This idealised model, also called normative model, is depicted in Figure 3. It is an aggregated version of the real model that shows only succession and flow between videos from weeks 1 to 6. The main reason for not showng all videos in a chain is the high number of videos in the MOOC. With over 60 videos, the model would not be readable in this paper. The model used in the experiment therefore specifies the first lecture in the series "Lecture 1.1: Data Science and Big Data (17 min.)" as the first task and the last lecture "Lecture 6.9: Data Science in Action (9 min.)" as the last task in the model.

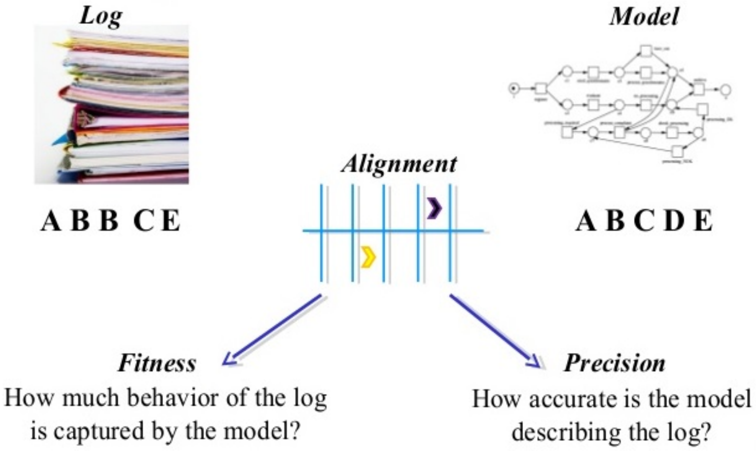

Log
A B B C E
Alignment
Model
A B C D E
Fitness
How much behavior of the log is captured by the model?
Precision
How accurate is the model describing the log?

Fig. 2: Alignment-based Conformance

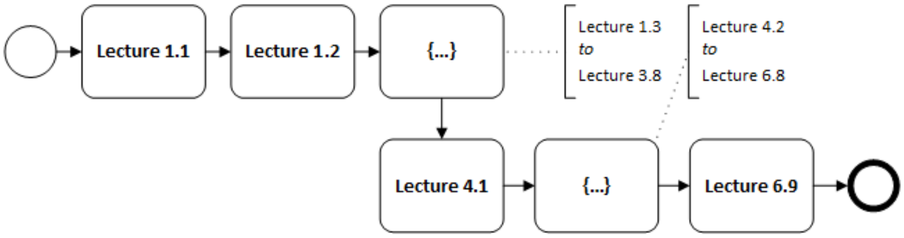


Fig. 3: BPMN Model for Sequential viewing of videos from Lecture 1.1 in Week 1 to Lecture 6.9 in Week 6

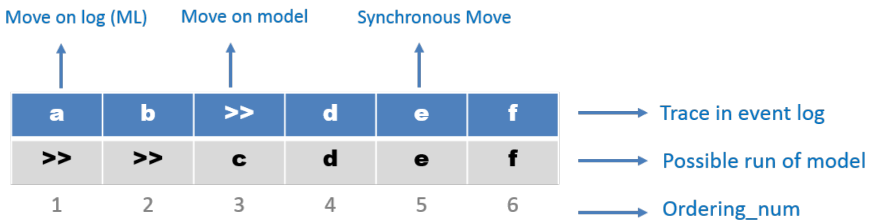


Fig. 4: Conformance Alignment moves

3 Defining Watch Status and Viewing Habit

There are are 2 critical aspects of students behavior that we model from the results of the conformance checking: watch status and viewing habit. With watch status, we aim at determining the sequence according to which each video is played, while theviewing habit defines the interval time between successive videos.

3.1 Video Watch Status

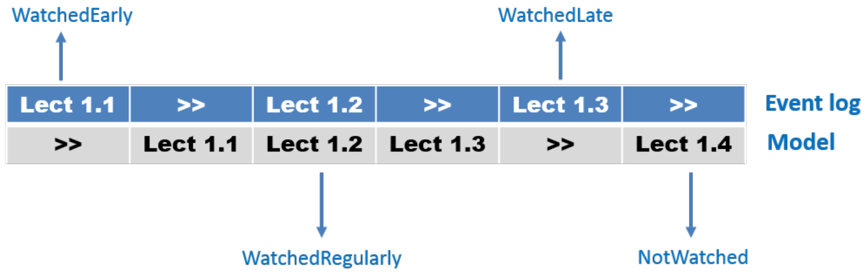
In order to label a video status, we consider moves that are generated by conformance alignment as seen in Figure 4. There are 3 types of moves that can be generated as a result. A move on log occurs when the task is found in the log only, a move on model occurs when the transition is only found in the model, and a synchronous move occurs when there is a match[2] . Hence, looking at these 3 moves, we define the watch status as follows:

```
SET Watch Status =
CASE WHEN move = ‘synchronous’ then ‘WatchedRegularly’
    WHEN move = ‘modelOnly’ then ‘NotWatched’
    WHEN move = ‘logOnly’ then
        CASE WHEN ordering_num in model < ordering_num in log
                then ‘WatchedEarly’
          ELSE ‘WatchedLate’
          END
END
```

As an illustration, we consider a possible run of log with 4 transitions (lectures): Lect1.1, Lect1.2, Lect1.3 and Lect1.4. We also consider an event log with trace $\langle$*Lect1.3, Lect1.2, Lect1.1*$\rangle$. With conformance alignments, we can identify the videos watch status as depicted in Figure 5 .

3.2 Viewing Habit

The viewing habit depends on the time at which each 2 successive videos are opened. There are numerous ways one can decide to label these intervals. In this



Lect 1.1	>>	Lect 1.2	>>	Lect 1.3	>>	Event log
>>	Lect 1.1	Lect 1.2	Lect 1.3	>>	Lect 1.4	Model

Fig. 5: Description of videos watch status

paper, we chose to count the number of minutes and define the thresholds as follows:

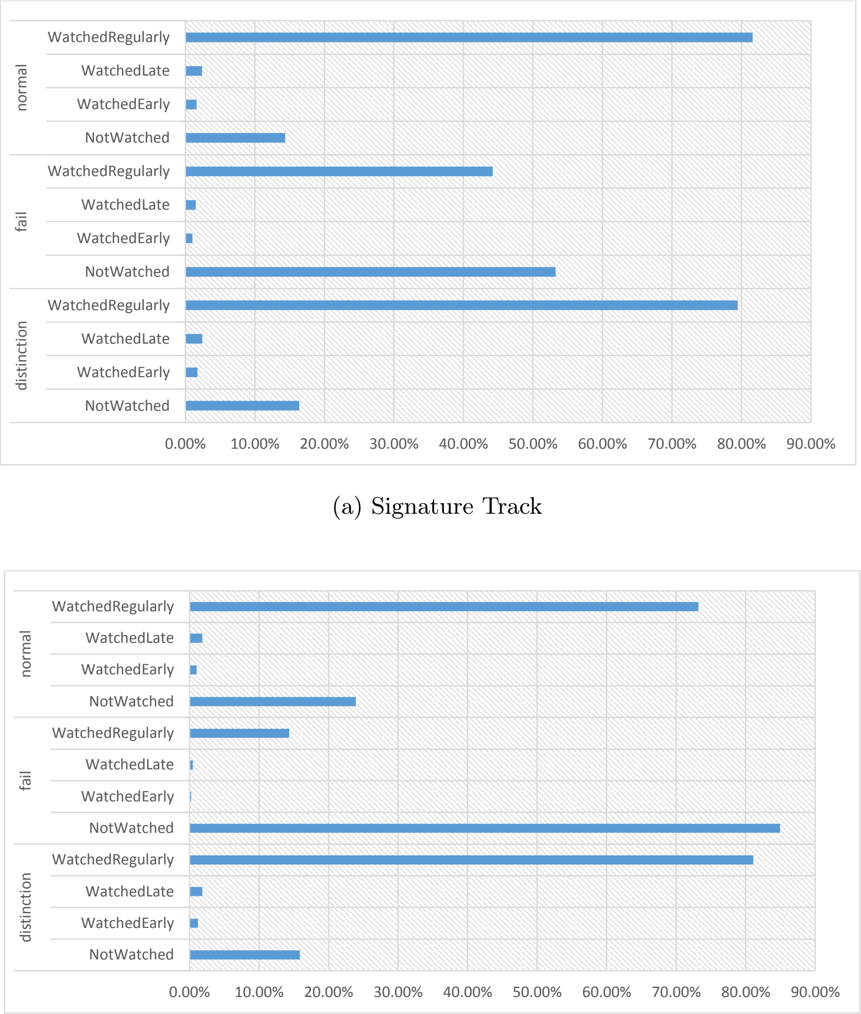
```
SET Viewing Habit =
CASE WHEN interval $\leq$ 30 then ‘InBatch’
          WHEN interval $\leq$ 60 then ‘After30min’
          WHEN interval $\leq$ 120 then ‘Hourly’
          WHEN interval $\leq$ 720 then ‘Halfdaily’
          WHEN interval $\leq$ 1440 then ‘Daily’
          WHEN interval $\leq$ 10080 then ‘Weekly’
          ELSE ‘Skipped’
END
```

4 Quantification of Learning Behavior

We performed conformance checking using the normative model in Figure 3 to analyze viewing behaviors for all subgroups of students. By exploring the alignment details we can then extract details about the overall watch status and viewing habits.

4.1 Watch Status

In Figure 6, a glimpse of the overall videos status for both signature and non-signature track students is given. We notice a confirmation that successful students appear more committed in watching videos than unsuccessful students. Furthermore, a demarcation can be observed between signature-track and non-signature track students as the former show more signs of committment in watching videos accross the different subgroups in comparison to their non-signature track counterparts. Successful students with the signature track certificate watched regularly in more than 80% of the videos as seen in Figure 6a, while those who failed watched 45% of the videos as seen in Figure 6b. The non-signature track successful students watched on average 80% regularly, while their failed students watched only 15% of the videos.



(a) Signature Track

(b) Non-Signature Track

Fig. 6: Overall watch status for the entire students population

We can break this information down into weeks. Over the period of 6 weeks, Figure 7 shows the watch status pattern accross the different weeks. We observe a commitment from the successful students from the first to the last week. While there is a progression for unsuccessful students in not watching videos from the 1st to the last week.

We can focus only on a few weeks, i.e. the 1st, 3rd and last, and get a better feel of how this behavior spans accross the different groups as demonstrated in Figure 8. The same can be done for every single week separately if needed.

Using the same details, we can visualize and quantify the status for each video either individually or according to the specific week it is was put online as seen in Figure 9.

It is also possible to visualize a correlation between a single video status and the corresponding weekly quiz as seen in Figure 10. Looking at Lecture 1.6, one can see the impact of its watch status on quiz 1. Students who do not watch this video incur an inferior average in comparison with the rest of the students who watched the video.

4.2 Viewing Habit

The viewing habit describes the time commitment in the students' learning behavior. Figure 11 indicates that in most part, successful students watch videos more in batch and do not mostly waste a lot of time between videos. Unsuccessful students on the contrary skip more videos than they watch.

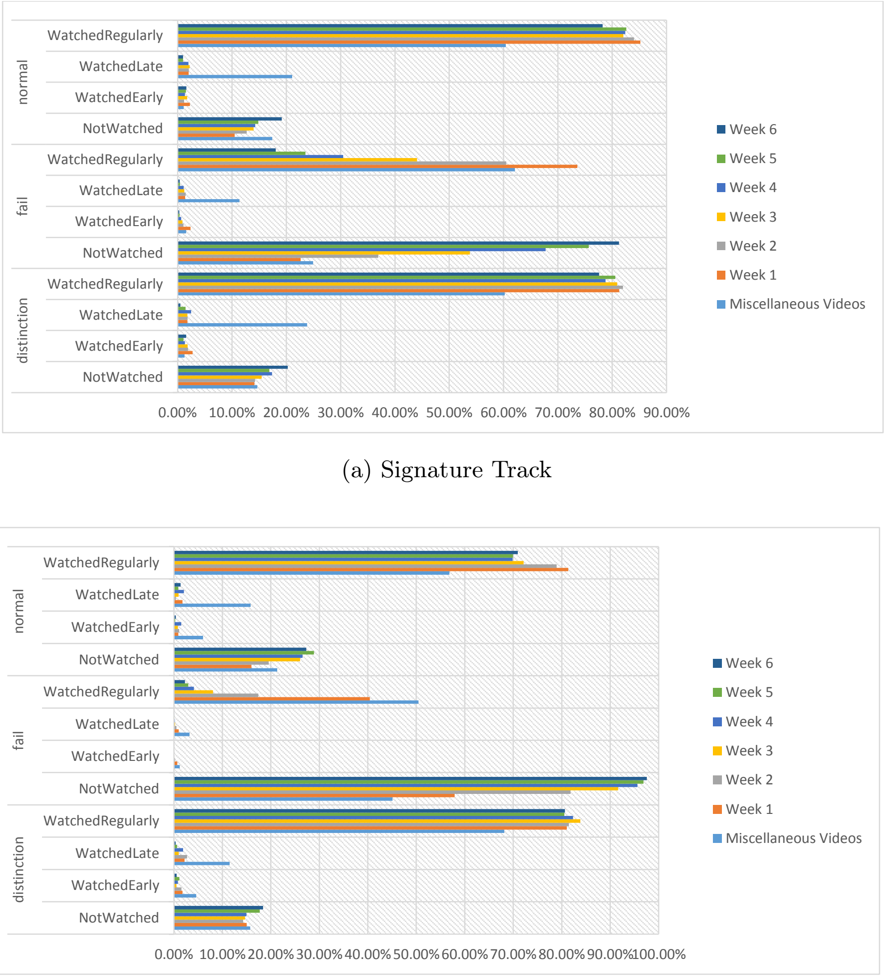
Breaking down this information, we can display a representation over 6 weeks as seen in Figure 12. There is a clear indication of the impact of viewing habit on performance and students' final grades. The most committed students, who watch mostly in batch appear to be more successful than the rest.

The opposite trend is observed with regards to usuccessful students who skip videos increasingly. As the MOOC starts, some of these students are devoted to watching but as time progresses, they stop watching certain videos and this shows accross the board for all unsuccessful students. Moreover, unsuccessful students' behavior pertaining to watching in batch progressively decreases as the weeks go by. The more videos were watched in batch in week 1, the less they are in week 6.

We can confirm this viewing trend by focusing on week 1 and 6 as seen in Figure 13. We can notice that for all students, the rythm of following lectures diminishes as weeks pass. Nevertheless, successful students watch successive videos in batch.

4.3 Viewing Habit vs. Watch Status

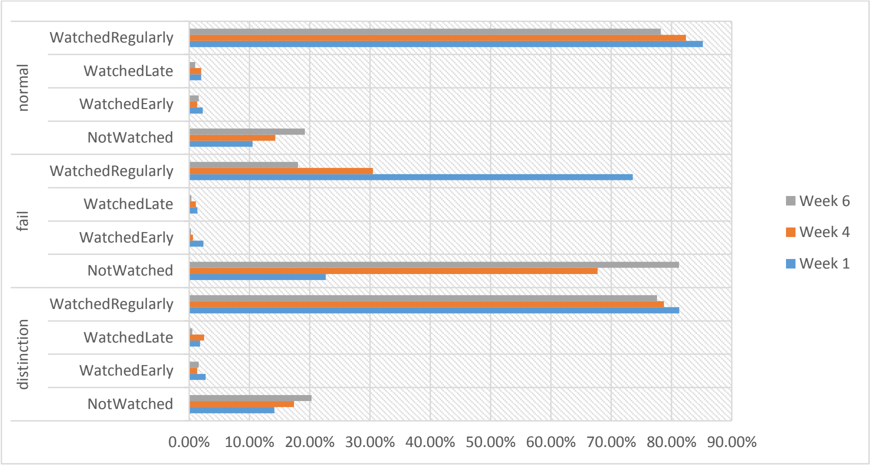
It is also interesting to visualize the correlation between viewing habit and watc status. Some of the questions we might try to answer are: “are students who watch videos in batch watching videos sequentially?”, “What are they skipping?”,



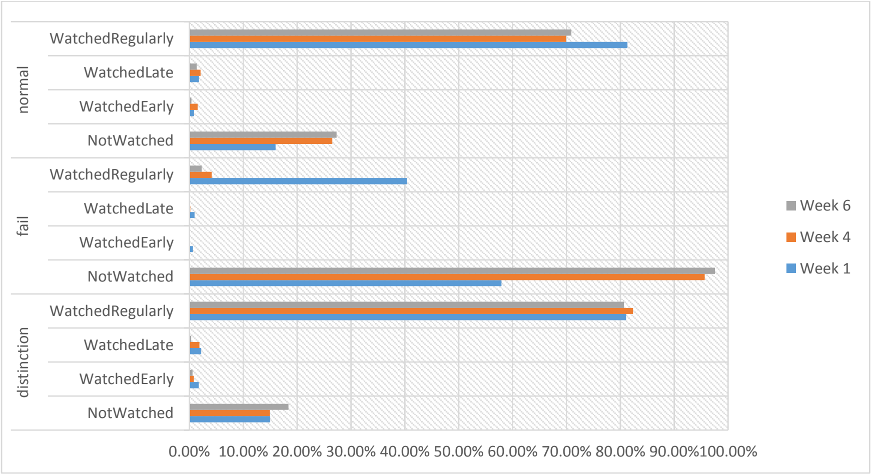
(a) Signature Track

(b) Non-Signature Track

Fig. 7: Overal watch status per week

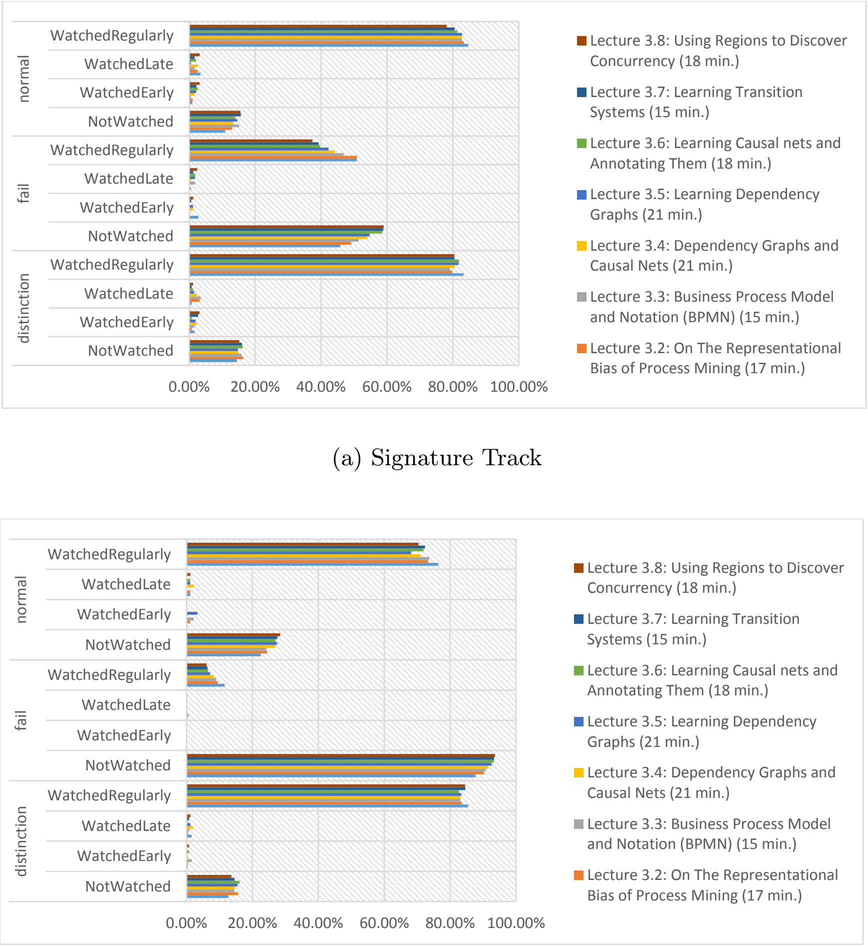


(a) Signature Track



(b) Non-Signature Track

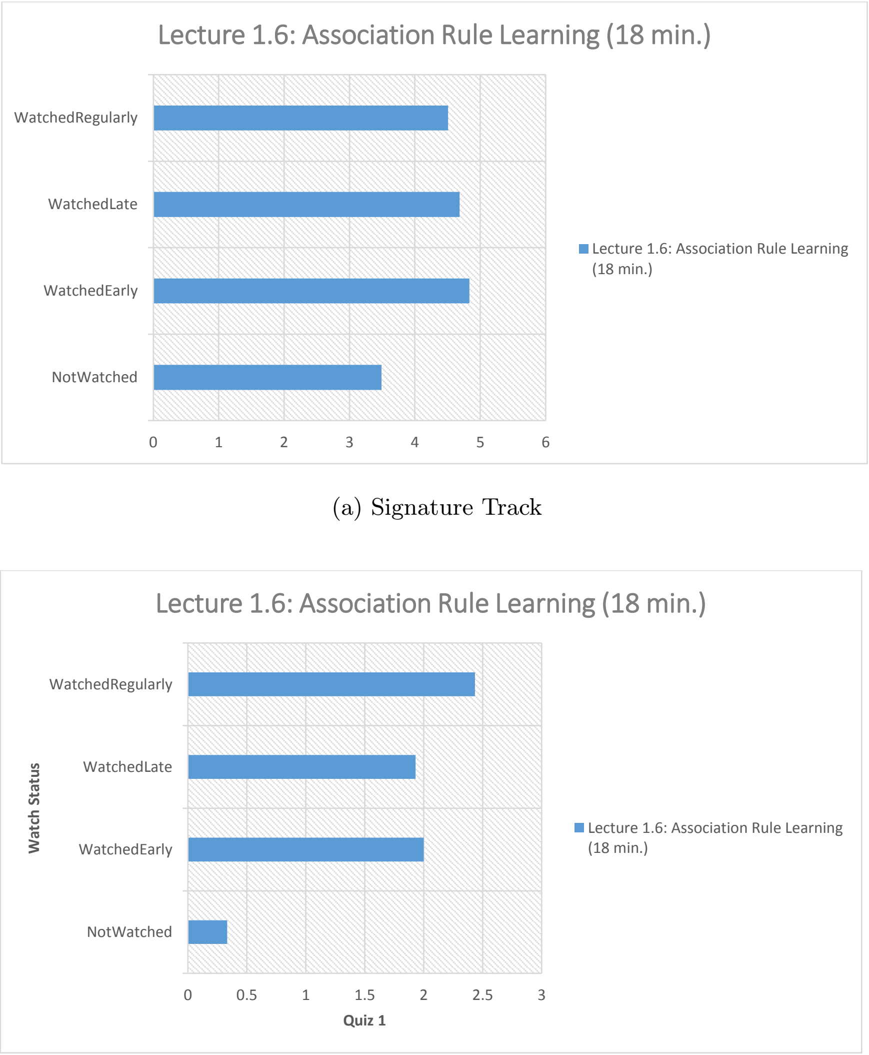
Fig. 8: Watch status for selected weeks



(a) Signature Track

(b) Non-Signature Track

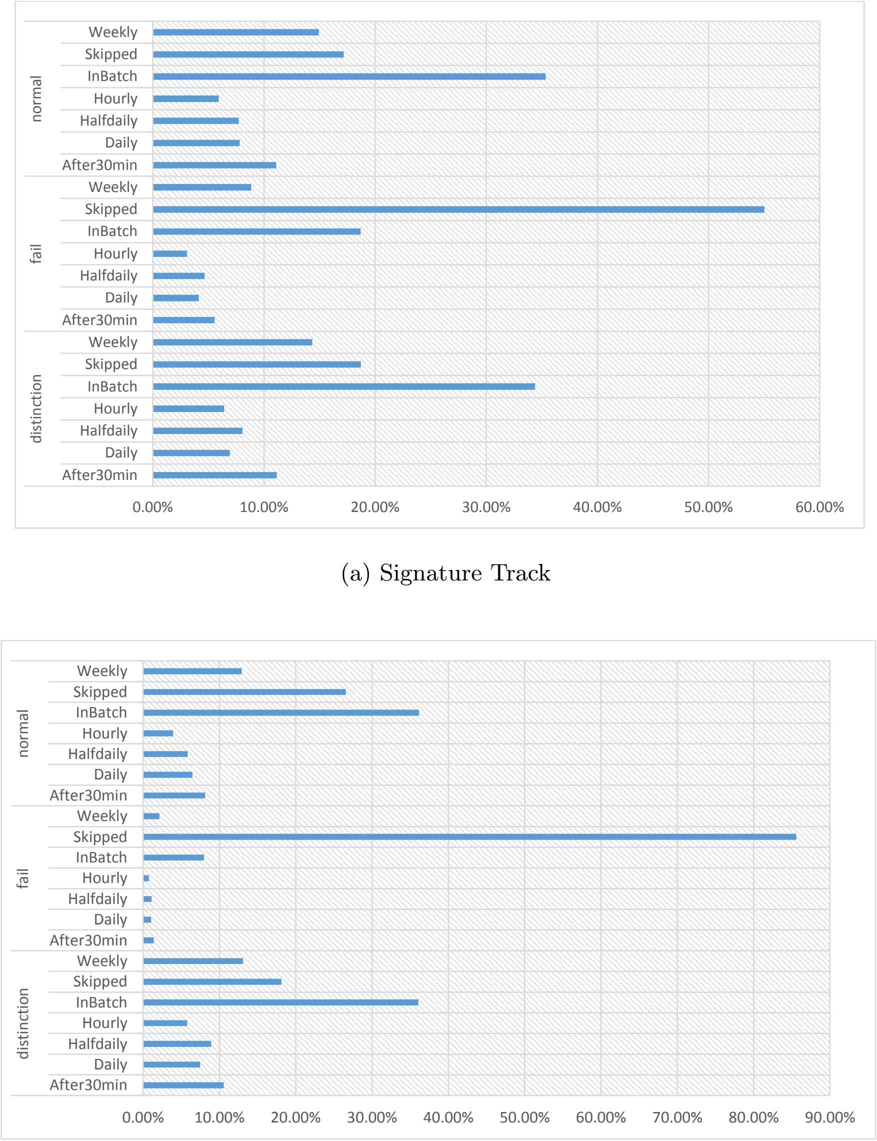
Fig. 9: Watch status for Lecture 3.2 to Lecture 3.8 in week 3videos



(a) Signature Track

(b) Non-Signature Track

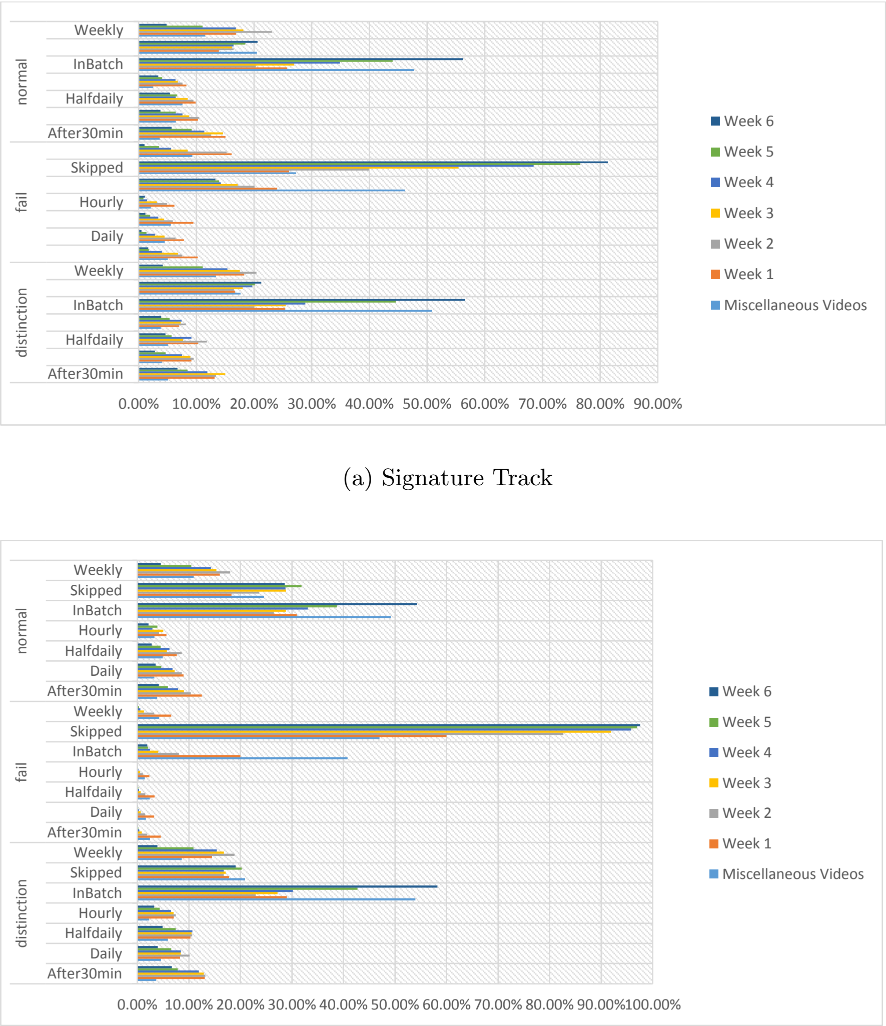
Fig. 10: A random lecture video in week 1 vs. Quiz 1



(a) Signature Track

(b) Non-Signature Track

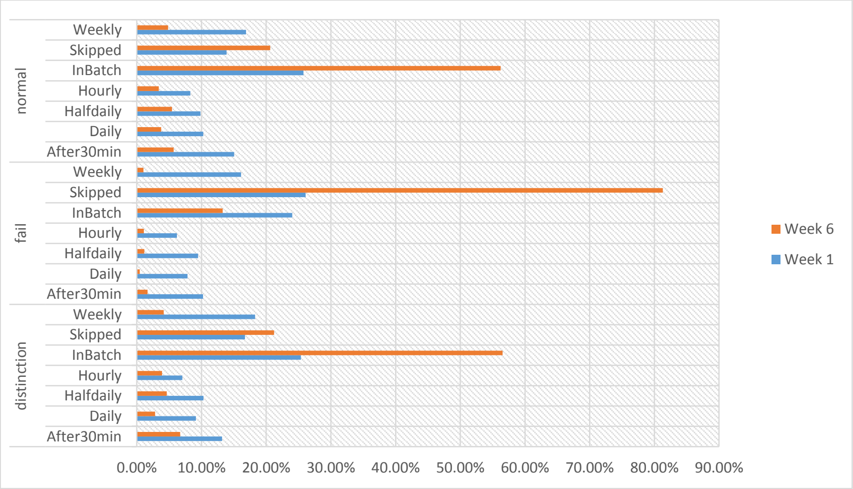
Fig. 11: An overall representation of viewing habit



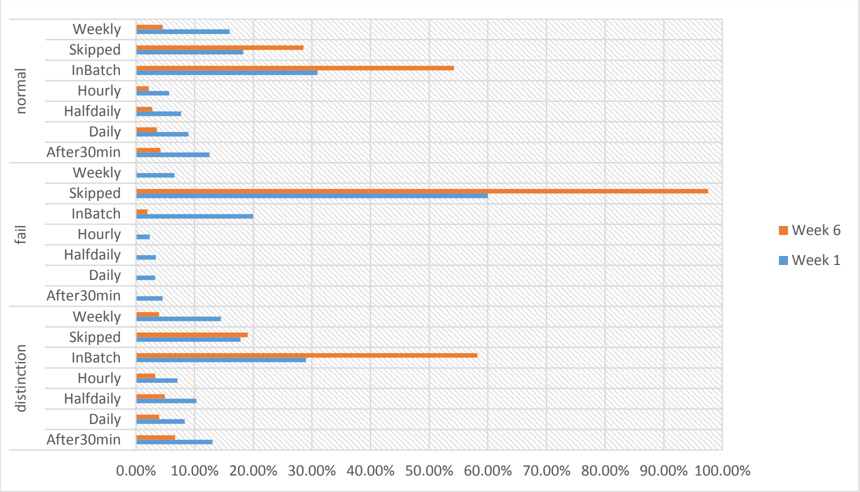
(a) Signature Track

(b) Non-Signature Track

Fig. 12: Viewing habits per week



(a) Signature Track



(b) Non-Signature Track

Fig. 13: Viewing habit in weeks 1 and 6

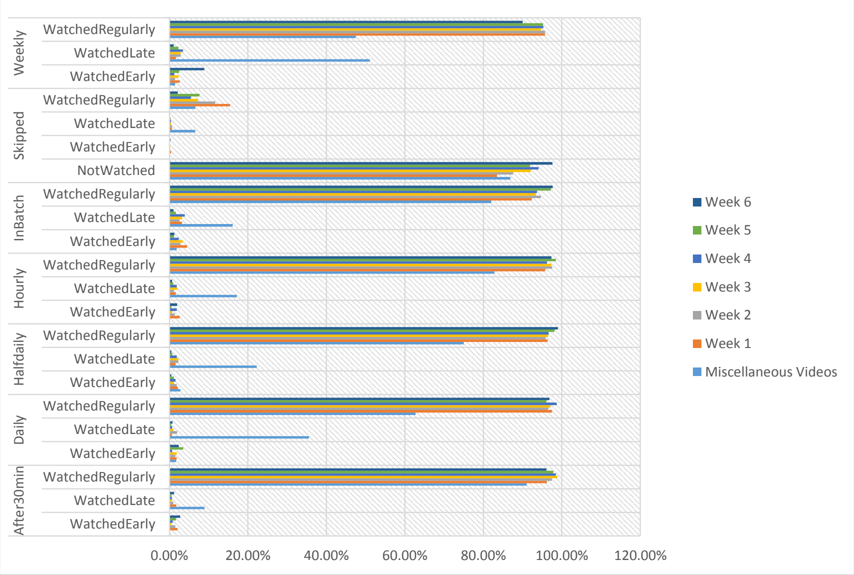


Fig. 14: Watch status versus viewing habit

"is there a link between both watch status and viewing habit and weekly performance?". In Figure 14, we observe that students who study in batch, mostly watch videos regularly (in sequence) than those who skip videos.

Figure 13 confirms that there exists a correlation between the way people watch videos and the interval of time between successive videos they watch.

5 Conclusion

It is certain that MOOCs are gaining considerable momentum and attracting interests from accross different professions. While many registered students seek to satisfy their curiosity, research suggest that many more harbour a genuine desire and need to foster their skills in any areas of interest. Allowing a lot of flexibility for professionals coupled with the reputation of some of the most prestigious institutions involved, MOOCs constitute a suitable learning solution even for people who are employed full-time.

The pressure and state of affairs call for continuous evaluation of the effectiveness of MOOCs, as many consider this learning approach to be in its infancy. With the emergence of the field of Learning analytics, there are proposals on means to study both the environment and students' involvement in MOOCs to spearhead improvements and appropriate interventions if any.

In this paper, we built on our initial work in [8] in demonstrating how process mining can be made use of in order to study students' behavior and its impact on their performance. We can highlight 2 major observations from our analysis: (1) students who do not waste time in watching videos and submitting quizzes tend to be more successful than those who follow the opposite path; (2) there are indications of the existence of a correlation between students' behavior and their final grade.

The advantages of using process mining for students' behavioral study is manifold. Primarily, it allows for the analysts to perform evidence-based fact-finding. This is because process mining works on real data based on students' activities online. Secondly, the techniques presented both in this paper and in [8] can provide critical insights that can lead to further types of analyses. For example, by locating people who do not study sequentially in earlier weeks, some directions could be provided to both advise and guide for a good achievemet level. Also, if a considerable number of students skip a lecture and this has no impact on their weekly or final quiz, this can trigger further investigations and measures.

In future, we aim at performing more experiments with other proces mining techniques. Moreover, we shall consider applying statistical significance tests in order to determine the levels of correlation between students' behaviors and their impact on their performances.

References

1. van der Aalst, W.M.P.: Process Mining - Discovery, Conformance and Enhancement of Business Processes. Springer (2011)
2. van der Aalst, W.M.P., Adriansyah, A., van Dongen, B.F.: Replaying history on process models for conformance checking and performance analysis. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2(2), 182–192 (2012)
3. Adriansyah, A., van Dongen, B.F., van der Aalst, W.M.: Conformance checking using cost-based fitness analysis. In: Enterprise Distributed Object Computing Conference (EDOC), 2011 15th IEEE International. pp. 55–64. IEEE (2011)
4. Adriansyah, A., Munoz-Gama, J., Carmona, J., van Dongen, B.F., van der Aalst, W.M.: Alignment based precision checking. In: Business Process Management Workshops. pp. 137–149. Springer (2013)
5. Christensen, G., Steinmetz, A., Alcorn, B., Bennett, A., Woods, D., Emanuel, E.J.: The mooc phenomenon: who takes massive open online courses and why? Available at SSRN 2350964 (2013)
6. Kay, J., Reimann, P., Diebold, E., Kummerfeld, B.: Moocs: So many learners, so much potential... IEEE Intelligent Systems (3), 70–77 (2013)
7. Liyanagunawardena, T.R., Adams, A.A., Williams, S.A.: Moocs: A systematic study of the published literature 2008-2012. The International Review of Research in Open and Distributed Learning 14(3), 202–227 (2013)
8. Mukala, P., Buijs, J., van der Aalst, W.: Exploring students' learning behaviour in moocs using process mining techniques. Tech. rep., Technische Universiteit Eindhoven (2015)
9. Munoz-Gama, J., Carmona, J.: A general framework for precision checking. International Journal of Innovative Computing, Information and Control (IJICIC) 8(7), 5317–5339 (2012)

10. Rozinat, A., van der Aalst, W.M.: Conformance checking of processes based on monitoring real behavior. Information Systems 33(1), 64–95 (2008)
11. Verbeek, H.M.W., Buijs, J.C.A.M., van Dongen, B.F., van der Aalst, W.M.P.: ProM 6: The process mining toolkit. In: Proc. of BPM Demonstration Track 2010. vol. 615, pp. 34–39. CEUR-WS.org (2010), http://ceur-ws.org/Vol-615/paper13.pdf