Resource usage analysis from a different perspective on MOOC dropout

Resource Usage Analysis from a Different Perspective on MOOC Dropout

Remi Brochenin[1], Joos Buijs[1], Mehrnoosh Vahdat[2], and Wil van der Aalst[1]

[1] Department of Mathematics and Computer Science, Eindhoven University of Technology, 5612AZ Eindhoven, The Netherlands,
[2] Department of Industrial Design, Eindhoven University of Technology, 5612AZ Eindhoven, The Netherlands
{r.brochenin, j.c.a.m.buijs, m.vahdat, w.m.p.v.d.aalst}@tue.nl

Abstract. We present a novel learning analytics approach, for analyzing the usage of resources in MOOCs. Our target stakeholders are the course designers who aim to evaluate their learning materials. In order to gain insight into the way educational resources are used, we view dropout behaviour in an atypical manner: Instead of using it as an indicator of failure, we use it as a mean to compute other features. For this purpose, we developed a prototype, called RUAF, that can be applied to the data format provided by FutureLearn. As a proof of concept, we perform a study by applying this tool to the interaction data of learners from four MOOCs. We also study the quality of our computations, by comparing them to existing process mining approaches. We present results that highlight patterns showing how learners use resources. We also show examples of practical conclusions a course designer may benefit from.

Keywords: MOOC, Learning Analytics, Educational Data Mining, FutureLearn, Dropout

1 Introduction

Learning Analytics (LA) and Educational Data Mining (EDM) use data to inform and support the stakeholders about the learning behaviour [2,3]. For instance, instructors can gain insight into the performance of learners, and learners can benefit from personalized guidance [4].

Analysis of interaction data of learners with Massive Open Online Course (MOOC) platforms is of growing interest for LA and EDM researchers. Indeed, the automatic collection and availability of data has raised interest in MOOCs [5] for gaining a better insight into properties of the learning behaviour. In this context, course designers are important stakeholders who are responsible for designing and planning courses. Understanding how the learners access and use resources, would help the course designers to adapt the learning materials to better fit the needs of learners.

Research on MOOC data often revolves around dropout behaviour and a notion of success [6]. This notion of success is based, for instance, on the completion of a proportion of the tasks or on the grade obtained in quizzes and

assignments. In this context, our work differs by relying on the fact that most of the audience of MOOCs is already highly educated, and many choose to study only parts of the course resources [7]. As a result, we do not direct our attention to the completion of the course. Indeed, a learner who does not complete the course, but still spends time viewing a variety of resources, will be considered as a valid learner. Just like a textbook reader who would need the information of only a couple of chapters, the MOOC learner can be selective about the course resources. Considering this observation, we choose to focus on the parts of the MOOC relevant for individual learners.

The main aim of this paper is to gain insight into the resource usage behaviour of MOOC learners, by adopting this view on dropout. For that purpose, we develop a prototype called RUAF (Resource Usage Analysis for FutureLearn) to derive features about each resource, reflecting how interesting this resource was for the whole audience, including those who did not complete the course. For instance, we determine how many learners come back to a resource for reference, or how many skip a resource. This prototype can be applied to any of the MOOC datasets collected by the FutureLearn platform. We made RUAF publicly available [8]. Finally, we confirm the use of our prototype by testing our approach over four MOOC datasets. We also compare our approach with the process mining method of alignments.

This paper is structured as follows. In Section 2 we present related work. Then, in Sections 3 and 4 we present the datasets studied in this paper and explain the RUAF prototype. We then demonstrate the application of our prototype on the datasets in Section 5. Finally, in Section 6 we conclude and indicate pointers for future work.

2 Related Work

In this section we give an overview of related work in LA and EDM specifically targetting MOOCs. After discussing the approaches most relevant to FutureLearn, we take a more detailed look at questions of understanding learner behaviour and then more precisely at questions about video usage.

2.1 FutureLearn: A Growing MOOC Platform

FutureLearn is a growing MOOC platform promoted by the Open University (UK). Course designers are provided with access to the interaction data of their learners. A variety of LA and EDM approaches have been applied to FutureLearn data [6]. For instance [9] provides some short insights into how FutureLearn, through developing analytics dashboards, tries to provide a better feedback to stakeholders.

However, the reported works insist on looking at data in the same way as traditional classrooms, in which the interactions between students are now electronic and hence logged. The summary in [6] is confirming our diagnostic, with in general a traditional view on dropout and the completion of a course, supporting a dichotomy between success and failure.

2.2 Analysis of Resource Usage in MOOCs

MOOC usage behaviour from the perspective of the resources has not been studied as often as the overall behaviour of individual learners. However, understanding how the MOOC materials are accessed can be valuable and helps to improve the quality of lessons and structure.

There has been some attempts at analyzing MOOC video interaction patterns to identify the problems in the resources and their difficulty level for learners. In [10] a clickstream analysis is described that is based on the available types of interactions with videos. This study focuses on perceived difficulty of videos for the learners and video revisiting behaviours, and provides insight for the course designers. For instance, they advise to reduce the information overload in the lecture slides so that the less strong students can follow the course better.

In [11], a visual analytical system is presented to help educators gain insight into the learning behaviour through clickstream data from Coursera. The study offers several types of visualizations that show the difference of behaviour while viewing videos of the course. Behaviours such as "pause" and "play" are measured every second of the course. This visualization highlights the parts of each video that are viewed more often, or where learners chose to pause. This informs the course designers about the parts of the videos learners are interested in. Similarly, in [12] the authors try to analyze video watching patterns through the detection of in-video dropout and peaks of activity within a video.

2.3 Learner Behaviour as a Process

In the context of LA and EDM, the event logs of learners can be considered as a temporal and ordered process. Some studies consider the interaction data of learners as process data and analyse for example whether the learners follow a planned curriculum.

For example, in [13], methods of process mining are applied and a framework is introduced to help educators analyze educational processes, and facilitate real-time detection of curriculum violations. Also, in [14], the authors quantify how well the learners follow the order of the curriculum, from the event log of the learner. They compute for instance a feature related to the frequency of events, as well as a delay between the availability of a resource and the access to said resource. Their purpose is to compare learner behaviours.

In [15] a more customized process mining approach was presented. An innovative measurement of the way students watch resources is introduced. A relation to each resource is computed for each student: the student can have watched the resource either on time, typically after the previous resource in the curriculum and before the next one, or early, or late. The tool is based on alignments, which we will describe later in this paper.

3 MOOC Data

In this section, we describe the way data is provided by FutureLearn since the aim of our prototype RUAF is to be used on data for any course using this MOOC

Table 1. An extract of event logs recorded by FutureLearn.

learner_id	resource	first_visited_at	last_completed_at
learner1	1.1	2016-07-11 00:02:28 UTC	2016-07-11 00:12:54 UTC
learner2	1.1	2016-07-11 00:20:30 UTC	2016-07-11 00:22:55 UTC
learner3	1.1	2016-07-11 00:34:18 UTC	2016-07-11 00:35:46 UTC
learner1	1.2	2016-07-11 00:38:20 UTC	2016-07-11 00:40:24 UTC

platform. We collected datasets through two courses offered by our university and we were granted access to two more datasets provided by external institutions.

3.1 Data Collection From FutureLearn

On the MOOC platform FutureLearn, each course is offered with a weekly basis. The weekly structure encourages the learners to follow a relatively linear approach to the course, since the weekly resources are provided to the learners once every seven days. Each week contains a list of resources numbered as 'week number'.'resource number' (e.g. 1.2 is the second resource in the first week). Each resource can be one of the following types: a video, an article, a discussion, a quiz, or an assessment-related item. The assessment-related items are tests similar to a quiz, or peer-reviewed assignments.

Courses may be provided through several runs, for each of which a separate dataset is provided by FutureLearn. The dataset contains varied data that describe the interactions of the learners with the platform and other learners. We focus here on the part of dataset that is called the 'step activity', which contains the temporal interaction data of learners with the MOOC resources and resembles an event log. Table 1 shows an extract of the event log by FutureLearn. For each learner and each resource the learner has accessed, there is an entry stating the first time the learner accessed that resource, and the last time they completed that resource. Note that compared to other platforms (e.g. Coursera), the data provided by the FutureLearn platform is less detailed.

3.2 Datasets

ProM course We collected data of a FutureLearn course called "Introduction to Process Mining with ProM", provided by Eindhoven University of Technology (TU/e).[3] This MOOC covers topics related to process mining and focuses on the practical use of the ProM tool. The duration of the course is four weeks and videos are the main resource type.

The first week is introductory, and the learners install and do basic analyses using ProM. The second and third weeks offer more advanced applications of ProM. Finally, the last week is studying a dataset through an assignment and discussions related to the assignment. For this week, no new topic is introduced, and no video is included.

[3] `http://www.futurelearn.com/courses/process-mining`

Additional datasets We obtained FutureLearn data from two more courses provided by the University of Twente. We used these datasets to further test our approach. These datasets are as follows.

Nano course: a FutureLearn course that is offered for a duration of four weeks, "Nanotechnology for Health: Innovative Designs for Medical Diagnosis".[4]

Ultra course: a FutureLearn course that is offered for a duration of six weeks, "Ultrasound Imaging: What Is Inside?".[5]

Converted course We obtained a larger fourth dataset based on the Coursera MOOC called "Process Mining – Data Science in Action" offered by TU/e. The duration of this course is six weeks. We chose this dataset so as to test the scalability of RUAF. We converted the detailed clickstream data provided by Coursera into the less informative and coarser format of the FutureLearn 'step activity' where only the first and the last access to each resource are kept.

4 RUAF: Resource Usage Analysis for FutureLearn

The aim of our work is to provide the MOOC designers with insight on the usage of the resources of their courses considering all participants. The novelty of our approach can be explained as follows.

Firstly, we note that [16] suggests that learners may participate in the course in their own preferred way while they may be classified as dropouts. These participants have their own pace and selection of the materials, and might not follow the structure of the course as planned by the course designer. In other words, MOOCs are different from a classroom where the students need to accomplish a certain percentage in the assessment. Learners are considered mature enough to choose to not follow the entire course, and only focus on sections of it. Hence, we do not consider dropout learners as course failures, and do not exclude them.

A second characteristic of our approach is the assumption, validated in the next section, that learners tend to view resources from the beginning up to a certain point. We call this point the 'dropout point' that refers to the resource after which the learner ceases following the course. We do not consider 'dropout point' as a negative term since knowing this point under our assumption means knowing the part of the course that interests a learner: from the beginning up until this point. We call our 'dropout point' assumption DPA.

This allows us to differentiate between two possible reasons a learner does not use a resource: (1) the learner has not passed their 'dropout point' and instead they are skipping the resource, (2) the learner has passed their 'dropout point' and is not active any more. The second case is not reported as skipping a resource.

We develop a prototype, RUAF, which studies the behaviour of each learner according to this view, and extracts a set of features for each resource, as described in the following sections. We made RUAF publicly available [8] for the

[4] `http://www.futurelearn.com/courses/nanotechnology-health`
[5] `http://www.futurelearn.com/courses/ultrasound-imaging`

research community. The architecture of this tool follows the structure presented in Figure 1.

4.1 Initial Preprocessing

We consider a learner has *done* a resource when they interact more than a certain threshold in spending time with that resource. The threshold in our analysis is set to one minute. Thus, we exclude any learner that has not spent more than one minute with any resource.

4.2 Dropout Point

The 'dropout point' is the basis for the rest of our computations. We aim to compute a reliable indicator of when each learner ceases to be interested in the course. For that purpose, we consider each learner independently from the others. The '*dropout point*' for each learner is defined as the earliest resource such that: the learner has done less than one third of the resources between that resource and any later resource. This definition makes use of the assumption DPA to attempt at capturing the earliest point such that the learner is not involved in the remaining part of the course.

We define 'dropout point' formally as follows. Let R be the total amount of resources in the studied course, and L be the set of learners. We define the function D which for any learner $l \in L$ and any two integers $1 \leq i < r \leq R$ returns $D(l, i, r)$, the number of resources that l has done between the $(i+1)^{\text{th}}$ resource and the r^{th} resource. We also define the property $P(i, l)$ for an integer $i \leq R$ and a learner $l \in L$, which holds if and only if for all r, if $i < r \leq R$ then $D(l, i, r) \leq (r - i)/3$. Finally, the 'dropout point' of a learner l is $\text{dropout}(l) = \min\{i, 1 \leq i \leq r \text{ and } P(i, l)\}$.

For instance, in a course with 9 resources, resource 3 is a good candidate for the 'dropout point' ($P(3, l)$ holds) if:

- the learner has not done resources 4 and 5 (otherwise it would be more than one third of them).
- the learner has done at most one of the resources 4, 5, 6, 7 and 8.
- the learner has done at most two of the resources 4, 5, 6, 7, 8 and 9.

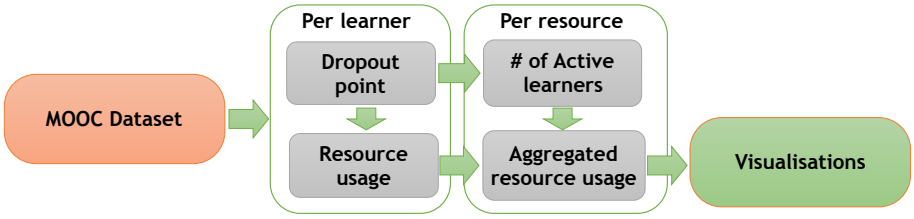

Fig. 1. RUAF prototype architecture.

We can then aggregate this feature for all learners at the resource level. We first compute for each resource how many learners are still active, by counting how many learners have not passed their 'dropout point' yet. In other words, we consider that learners are active until their dropout point. In mathematical terms, for the r^{th} resource, the '*active*' feature is the number of $l \in L$ such that $\text{dropout}(l) \geq r$.

We compute the proportion of active learners who had their 'dropout point' at a particular resource, and obtain the feature we call '*drop*'. In more formal terms, for the r^{th} resource, 'drop' is the number of $l \in L$ such that $\text{dropout}(l) = r$ divided by the 'active' feature of the r^{th} resource.

The way we obtain the 'dropout point' allows us to accurately exclude those learners who are not interested in the course from a certain point while keeping those who are interested but they are selective in using the resources.

4.3 Usage Features

From the 'dropout point' we can determine for each learner and resource two features. First, given a learner and a resource they overlooked, we can know whether it was a skipped resource, or the learner dropped out before the resource. We then aggregate at the resource level and divide that total by the number of learners still active (the 'active' feature of the resource), so as to obtain the '*skip*' feature: the proportion of active learners that have not done that resource. Then, similarly, given a learner and a resource they have done, we can know whether the learner was simply active, or the learner peeked at a resource while having already passed the 'dropout point'. Then by aggregating at the resource level and dividing by the total number of learners, we obtain the feature about learners that '*peek*' at a given resource (opposite of 'skip').

We also compute features relevant to which order learners use to view resources. First, we choose a threshold k (two in our computations) for the minimum number of resources that label the resource as done late or early. We say that a resource r is don late if at least k resources were done before r while they should have been done after r according to the curriculum; and these k resources are not good candidates at being done early w.r.t. r. Formally, given a learner, we define two functions on resource r seen by this learner. Let $A(r)$ be the set of resources that appear before r in the curriculum, but that the learner has started after starting r. Let $B(r)$ be the set of resources that appear after r in the curriculum, but that the learner has started before starting r. We say that:

- r was seen early: If there are at least k resources $r_1, \ldots, r_k$ such that for all r_i, both $r_i \in A(r)$ and the size of $B(r_i)$ is smaller than the size of $A(r)$;
- r was seen late: If there are at least k resources $r_1, \ldots, r_k$ such that for all r_i, both $r_i \in B(r)$ and the size of $A(r_i)$ is smaller than the size of $B(r)$;
- r was seen on time: Otherwise.

We then aggregate these notions at the resource level. We count for each resource how many learners saw it late or early, divide the obtained figure by the number of active learners, and obtain the features '*late*' and '*early*'.

Table 2. Extracted features of a resource, and their description.

feature	**description**
'active'	total number of learners who have not passed their 'dropout point' yet
'drop'	proportion of active learners in their 'dropout point' at a particular resource
'skip'	proportion of active learners who skip a particular resource
'peek'	proportion of learners who have done a particular resource while they have already passed their 'dropout point'
'early'	proportion of active learners who have done a particular resource early
'late'	proportion of active learners who have done a particular resource late
'back'	proportion of active learners who come back to a particular resource

Finally, we compute whether a learner revisited a resource, for instance for review or reference. We count how many resources that appear later in the curriculum were visited for the first time between the moment a particular resource was visited for the first time and for the last time. If the number of resources is more than a chosen threshold (which we call the *coming back threshold*, and set to three in our study), we say that the learner came back to this resource. Then we aggregate this at the resource level, counting for each resource how many learners came back to it, and we divide this figure by the number of active learners, and obtain the feature '*back*'.

A summary of extracted features is presented in Table 2.

4.4 Alignments

Computing whether a resource is done late or early, as well as computing the 'dropout point', can be obtained through process mining techniques, with the help of alignments. This method compares a process model with an event log, and verifies if they match [17]. In [15], a measure is introduced for determining if a resource is done late or early with respect to an expected process model. We extended their proposed method to be able to also compute the 'dropout point'.

We create a process model (Petri net) representing the order of resources set by the course designer as in [15], with a modification. For each place situated after a transition r of the Petri net defined in [15], we add a transition drop-r from that place to the end place. This allows us to measure the dropout behaviour by computing alignments to this model.

We applied this modified tool to be able to compare our results to [15].

4.5 Parameters

RUAF can be tuned to the data of a particular course with these parameters.

Minimum of time spent on a resource: we consider that any learner who spends less than one minute on a resource is the same as a learner who has spent no time on that resource (and hence have not done that resource). This time limit can be modified to any duration.

Dropout threshold: the proportion of resources that a learner has not done after the 'dropout point'. We set this threshold to one third, but this can be modified to any value. It can be useful to set a lower threshold if the course contains multiple resources that are non-mandatory, such as discussions, or links to further information.

Coming back threshold: the threshold used to compute the 'back' feature, set to three in our analysis.

Early and late threshold: the threshold for being late or early (k in the definition) is set to two in our analysis.

5 Results

In this section we present the results from applying RUAF to four datasets, and compare our results to the method of alignments in [15].[6]

5.1 Application of RUAF to Four MOOC Datasets

Here we present the results of the resource usage analysis with our prototype RUAF for the four MOOC datasets introduced.

We only consider the learners who spent more than one minute on any single resource.With this view: the *ProM course* had 908 learners, the *Nano course* had 935 learners, the *Ultra course* had 2384 learners and the *Converted course* had 12026 learners.

Drop We first study the 'drop' feature, which can be seen as the dropout rate per resource. According to literature, such as [18], we expected to see a much larger dropout rate at the beginning of each course, decreasing throughout the course. This expectation is confirmed in the ProM course, the Nano course and the Converted course. However, the Ultra course disproves this hypothesis. The 'drop' feature is relatively stable throughout the course, with a few outliers.

In all the courses, an interesting pattern of 'drop' emerges: at each transition from one week to the next, a small peak of dropout occurs. It may be slightly before the end of the week or slightly after the beginning of the next week, but is remarkably systematic. This pattern is the most notable in the Ultra course, see Figure 2 (note that as in all figures of the article, lighter bars correspond to resources of even weeks). The outliers of the Ultra course are all very close to a change of week, with the exception of resources number 5 and 36.

An analysis of outliers in the dropout pattern shows a lack of interest in assignments for a very large part of the learners. For instance, a notable outlier is the last week of the ProM course, during which the dropout rate suddenly increases. This week exclusively contains an assignment and discussions, resulting in only a third of the active learners to remain.

[6] Refer to Appendix A for all the visualizations provided by RUAF applied to the four MOOC datasets, as well as a more detailed report of the results from the method of alignments.

Skip and Peek The feature 'skip' can be seen as the proportion of learners still active at a particular resource who chose not to pay attention to that resource. The proportion of active learners who 'skip' the video materials is quite low.[7] Figure 3 shows the 'skip' behaviour of the ProM course for all the materials. The non-video materials (all long bars in Figure 3 are non-video), and particularly articles, tend to be skipped much more than video resources. The very high 'skip' rate for articles can be explained as many are not mandatory. For the videos, excluding three outliers, 'skip' is remarkably stable around 10% for the first two weeks, then decreases by half for the third week, while still being stable. The last week has no video. As a comparison, in the Converted course, 'skip' is stable around 20% for videos throughout the course.

The 'peek' behaviour shows a stable and very low rate for all resources, with few exceptions. This observation is visible in all four courses, for instance in the Nano course (see Figure 4) the 'peek' varies from 0% to 2% for the majority of

[7] In the case of the ProM and Converted courses, for which we know the resource types.

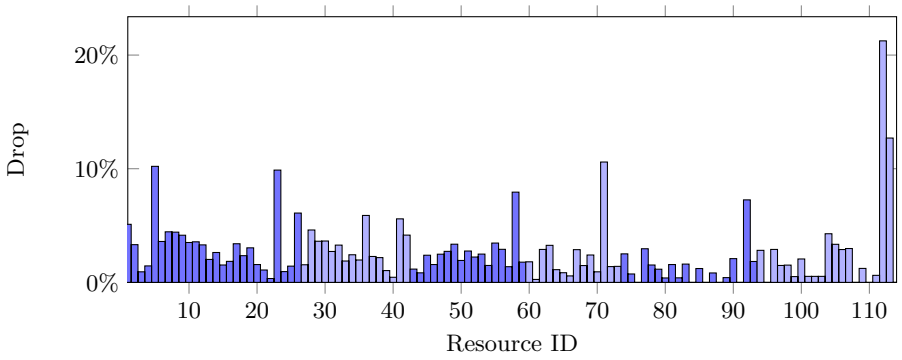

Fig. 2. 'Drop' for each resource in the Ultra course

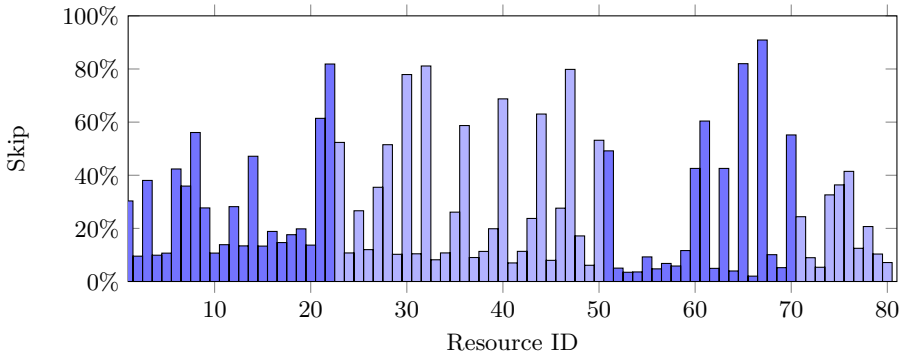

Fig. 3. 'Skip' for each resource in the ProM course

resources. The outliers (all above 4%) can be interesting for the designers of this course. These resources may be an indicator of specific topics that the learners who dropped out earlier had a particular interest in.

Our analysis based on 'skip' and 'peek' features confirms the hypothesis that learners access and use most resources until a 'dropout point'. Indeed, the learners have done on average about 90% of videos up until their 'dropout point', and perform nearly no action after the computed 'dropout point'. This also indicates that our computation of 'dropout point' is meaningful.

Early, Late and Back The values for 'early', as well as 'late', are very low in all courses, except the Ultra course. The Ultra course, shown in Figure 5, is characterized by higher values of 'early' (generally above 1% from the second week) compared to the other courses. Also, the high-valued outliers for the 'early' feature are generally at the beginning and end of the weeks. This may reflect the desire of learners to know what remains in the course.

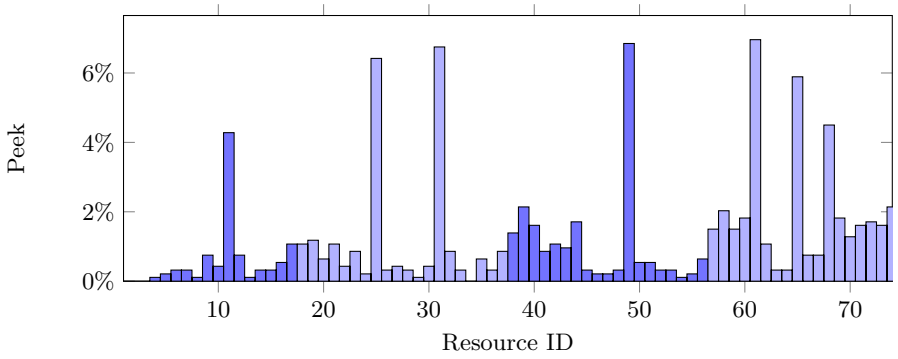

Fig. 4. 'Peek' for each resource in the Nano course

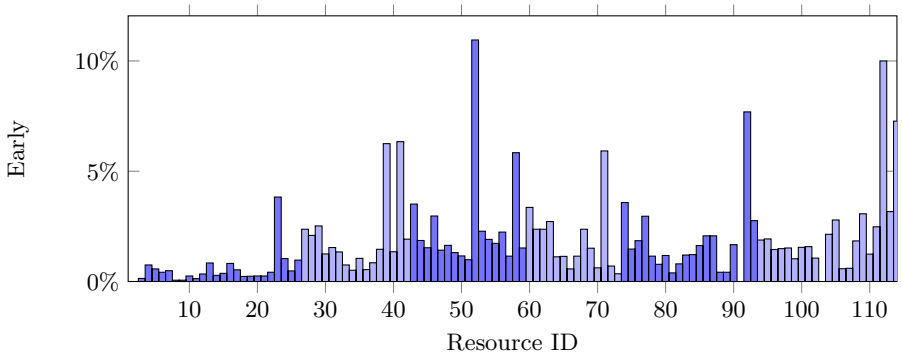

Fig. 5. 'Early' for each resource in the Ultra course

In all courses, 'back' is relatively stable, characterized by a decrease at the end of each week. The ProM course, in Figure 6, is the only one exhibiting a different behaviour. In its last week, the values of 'back' sharply increase. This can be explained by learners trying to find answers to their final assignment by accessing to the discussions at the end of the week.

5.2 Comparison with Alignments

We compare our results with the work of [15], which uses process mining to compute when a learner sees a resource early or late. All the aggregated features we obtained from the alignments show the same patterns as those from RUAF, which emphasizes that we are computing the same notion. Considering individual learners, we obtain results that are identical for the large majority of the cases. There are some differences, most of which are explained by two factors.

Firstly, with the alignment method from [15], as long as two items are switched[8], one of them (chosen arbitrarily) will be early or late. On the other hand, we request more than two items not appearing in the expected order so as not to need arbitrary choices.[9]

Second, the 'dropout point' is generally chosen earlier by alignments, since the cost of handling unordered events may be higher than just considering that the end of the process has been reached.[10] This also leads to a difference in the computation of an aggregated 'drop' from the alignment results. There are about 20% less learners 'active' per resource with this computation, and on average learners have their 'dropout point' more than five resources earlier.

As a conclusion, although in the examples studied the obtained patterns in the aggregated data are very similar, our approach has better results than

[8] Such as 3 and 4 in 1–2–4–3–5–6.

[9] For instance, in 1–2–4–3–5–6, it is unclear which of 3 or 4 is abnormal. Whereas in 1–4–2–3–5–6, resource 4 is seen early.

[10] As an example, if the trace is 1–2–6–5–4, getting to resource 6 before dropout in an alignment would have a higher cost than choosing a dropout at resource 2.

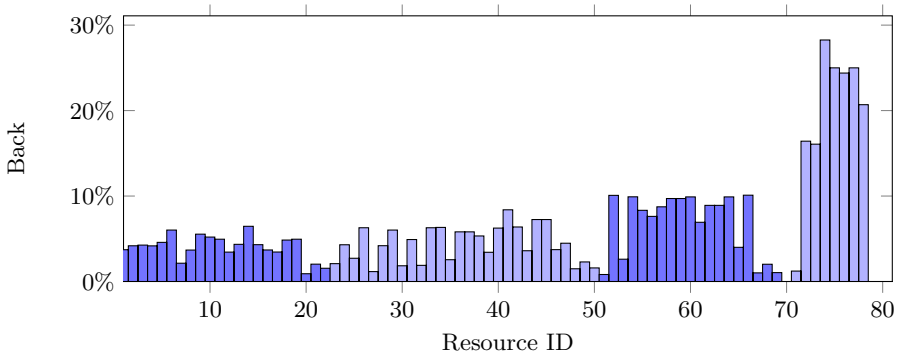

Fig. 6. 'Back' for each resource in the ProM course

alignments for providing process-related insight into the data we study. First, our computations never choose arbitrarily one possibility over another. Second, having a complex non-linear behaviour does not encourage the algorithm to set the 'dropout point' earlier.

6 Discussion and Conclusions

In the context of LA and EDM, there is a strong potential for progress in the feedback systems of MOOC platforms [6]. Developing ways to automatically analyze resource usage and generate visualizations can be a valuable tool for course designers [10,11]. Course designers want their content to reach a large audience, this is not limited to people following the whole course. From this perspective, learners who use just a part of the course are as important as the certificate earners. In this context, our work is a step towards considering all the users equally when providing insight into the resource usage.

We developed a prototype, called RUAF, that measures the resource usage properties through a novel approach toward 'dropout point'. We made RUAF publicly available [8] for the research community. We also studied the quality of our computations, in particular by comparing them to what process mining can compute and to the current statistics provided by FutureLearn platform. We showed that our approach provides better results compared to alignments computation of [15], while having much simpler semantics.

We finally showed in Section 5 some examples of practical conclusions a course designer may draw thanks to our prototype. For instance, our prototype allows to detect which learning resources are skipped or revisited by the learners, which resources provoke dropout, which type of materials (articles, videos, assignment, etc.) are more attractive for the learners, and which ones are accessed earlier or later than planned.

Such feedback is valuable for the course designers to tune the resources to the needs of learners and direct their time and effort to the parts that need more attention. For instance, if the order of resources in the curriculum is not followed by learners, they can modify the order to have a more balanced curriculum that is easier to follow. Also, detecting outliers in 'skip', 'drop', and 'back' can help them to exclude the resources that are problematic for the learners, change their type, or reduce their difficulty level.

In the future, our approach can be implemented in the FutureLearn platform to automatically recognize the resource usage properties, and provide feedback and recommendations to the course designers. Additionally, we aim to extend our approach to analyze more properties of resource usage in particular in the case of MOOC platforms that provide more detailed user interaction data. For instance, we can have a more detailed view on how students come back to a resource, how often they revisit resources and how they transit between the resources. Finally, with very detailed data on video views (e.g. play and pause events), one could be able to pinpoint particular patterns of usage within a video. For instance, within a video there could be particular moments that provoke going back to

another resource, or skipping to the next resource. Course designers would then be able to relate such information with the way a video is built, or with topic changes within a video.

References

1. Buckingham Shum, S.: Learning analytics. UNESCO Policy Brief (2012)
2. Chatti, M.A., Dyckhoff, A.L., Schroeder, U., Thüs, H.: A reference model for learning analytics. International Journal of Technology Enhanced Learning **4**(5) (2012) 318–331
3. Vahdat, M., Ghio, A., Oneto, L., Anguita, D., Funk, M., Rauterberg, M.: Advances in learning analytics and educational data mining. In: Artificial Neural Networks, Computational Intelligence and Machine Learning. (2015)
4. Papamitsiou, Z., Economides, A.: Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Educational Technology & Society **17**(4) (2014) 49–64
5. Baker, R., Yacef, K.: The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining **1**(1) (2009) 3–17
6. Ferguson, R., Coughlan, T., Herodotou, C., Scanlon, E.: MOOCs: What the research of FutureLearn's UK partners tells us. The Open University (2017)
7. Yuan, L., Powell, S.: Moocs and open education: Implications for higher education. `http://publications.cetis.org.uk/2013/667` (2013)
8. Brochenin, R.: Ruaf. `http://github.com/brochenin/RUAF` (2017)
9. Chitsaz, M., Vigentini, L., Clayphan, A.: Toward the development of a dynamic dashboard for futurelearn moocs: Insights and directions. In: ASCILITE. (2016)
10. Li, N., Kidziński, Ł., Jermann, P., Dillenbourg, P.: MOOC video interaction patterns: What do they tell us? In: European Conference on Technology Enhanced Learning. (2015)
11. Shi, C., Fu, S., Chen, Q., Qu, H.: Vismooc: Visualizing video clickstream data from massive open online courses. In: Visualization Symposium. (2015)
12. Kim, J., Guo, P.J., Seaton, D.T., Mitros, P., Gajos, K. Z .and Miller, R.C.: Understanding in-video dropouts and interaction peaks in online lecture videos. In: Learning @ Scale. (2014)
13. Trcka, N., Pechenizkiy, M.: From local patterns to global models: Towards domain driven educational process mining. In: Intelligent Systems Design and Applications. (2009)
14. Boroujeni, M.S., Sharma, K., Kidziński, Ł., Lucignano, L., Dillenbourg, P.: How to quantify student's regularity? In: European Conference on Technology Enhanced Learning. (2016)
15. Mukala, P., Buijs, J.C.A.M., Leemans, M., van der Aalst, W.M.P.: Learning analytics on coursera event data: A process mining approach. In: Data-driven Process Discovery and Analysis (SIMPDA). (2015)
16. Onah, D.F., Sinclair, J., Boyatt, R.: Dropout rates of massive open online courses: Behavioural patterns. In: EDULEARN. (2014)
17. van der Aalst, W.M.P.: Process mining: Data science in action. Springer (2016)
18. Breslow, L., Pritchard, D.E., DeBoer, J., Stump, G.S., Ho, A.D., Seaton, D.T.: Studying learning in the worldwide classroom: Research into edX's first MOOC. Research & Practice in Assessment **8**(Summer 2013) (2013) 13–25

A All visualisations

We present here all visualisations of resource usage analysis obtained through the work that led to this aticle. The subsections A.1, A.2, A.3, and A.4 are the visualisations of RUAF, exactly as the tool provides them. The subsections A.5 and A.6 present features that are obtained through alignements.

A.1 Features for the ProM course

Figure 7 contains the graph of the number of learners that are still active at each resource.

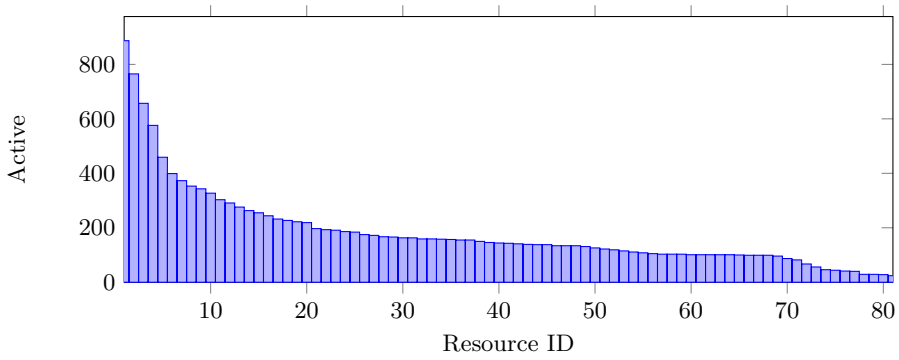

Fig. 7. Active learners for each resource in the ProM course

Figures 8, 9, 10, 11, and 12 contain respectively the graphs of the features 'drop', 'skip', 'back', 'early', and 'late' for each resource. They are features computed as a proportion of the number of active learners. Figure 13 contains the graph of the feature 'peek', which is computed as a proportion of all learners. In these six figures, the change of week is shown by changing the shade of blue. The even weeks are lighter than the odd weeks.

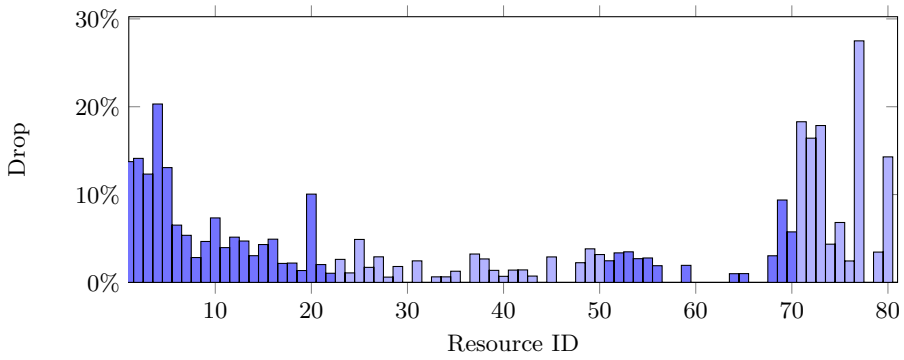

Fig. 8. 'Drop' for each resource in the ProM course

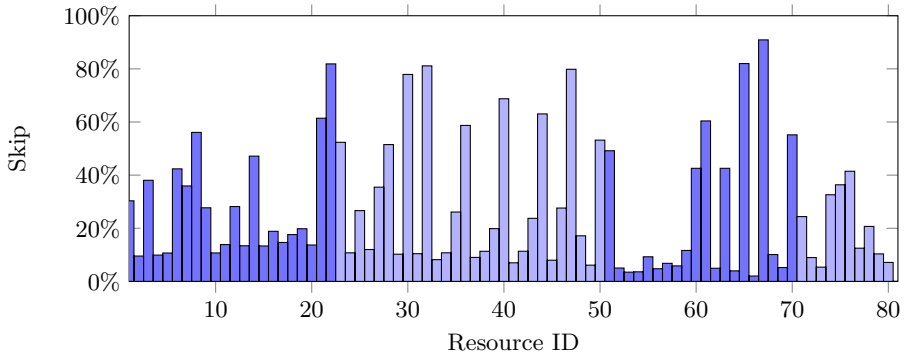

Fig. 9. 'Skip' for each resource in the ProM course

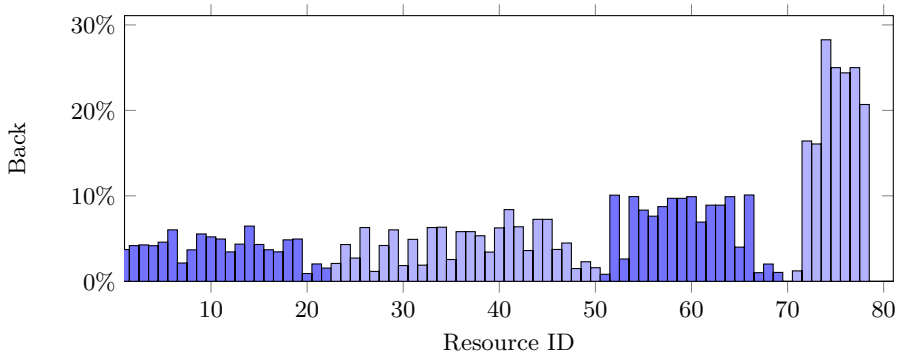

Fig. 10. 'Back' for each resource in the ProM course

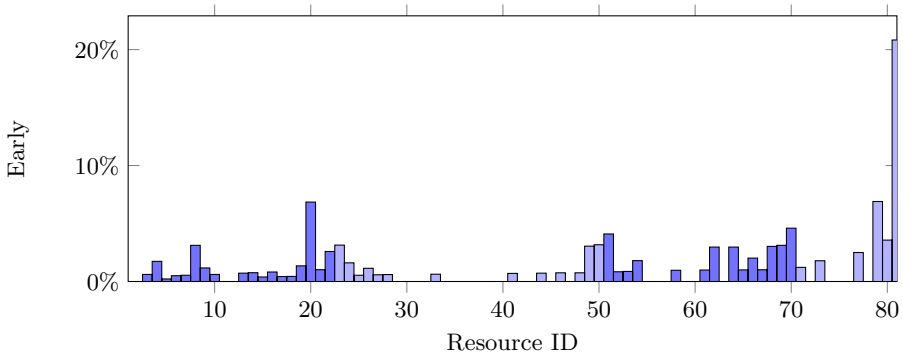

Fig. 11. 'Early' for each resource in the ProM course

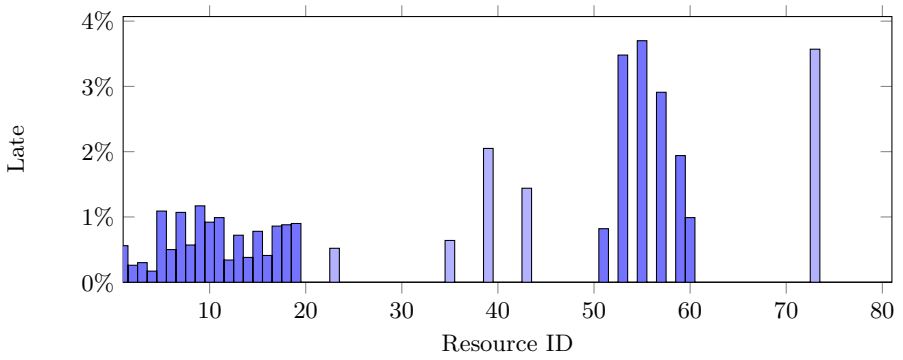

Fig. 12. 'Late' for each resource in the ProM course

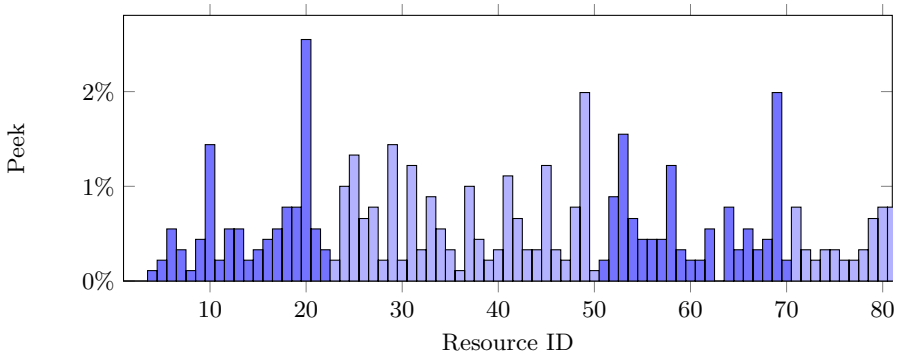

Fig. 13. 'Peek' for each resource in the ProM course

A.2 Features for the Nano course

Figure 14 contains the graph of the number of learners that are still active at each resource.

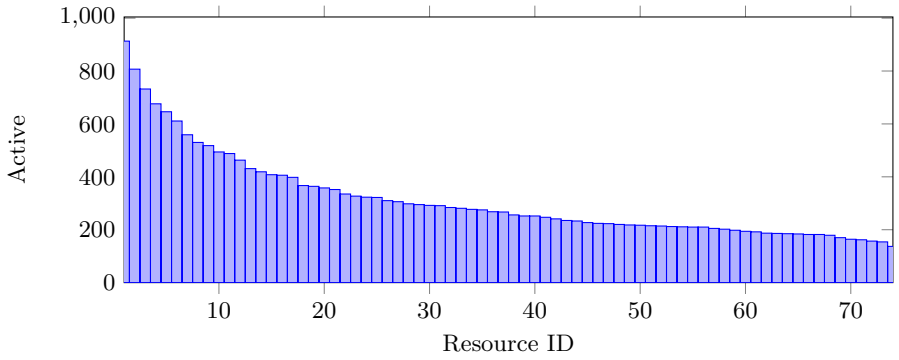

Fig. 14. Active learners for each resource in the Nano course

Figures 15, 16, 17, 18, and 19 contain respectively the graphs of the features 'drop', 'skip', 'back', 'early', and 'late' for each resource. They are features computed as a proportion of the number of active learners. Figure 20 contains the graph of the feature 'peek', which is computed as a proportion of all learners. In these six figures, the change of week is shown by changing the shade of blue. The even weeks are lighter than the odd weeks.

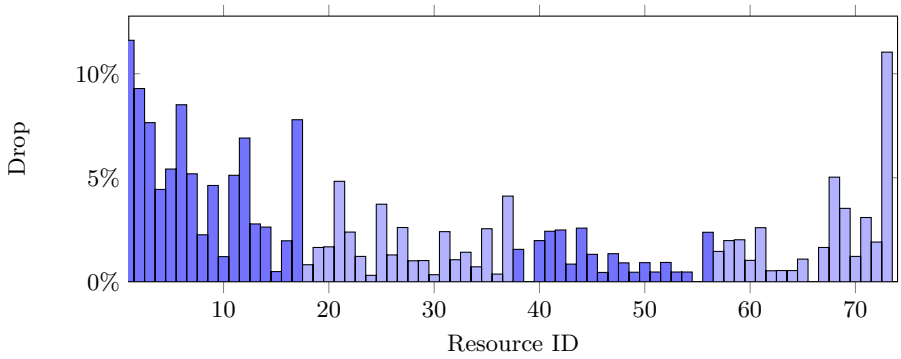

Fig. 15. 'Drop' for each resource in the Nano course

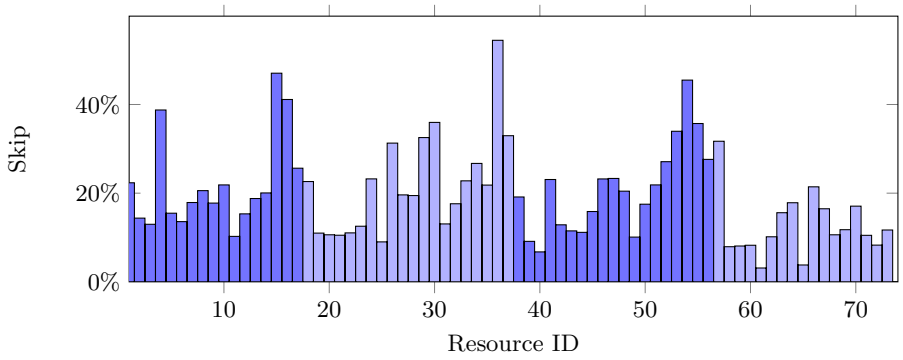

Fig. 16. 'Skip' for each resource in the Nano course

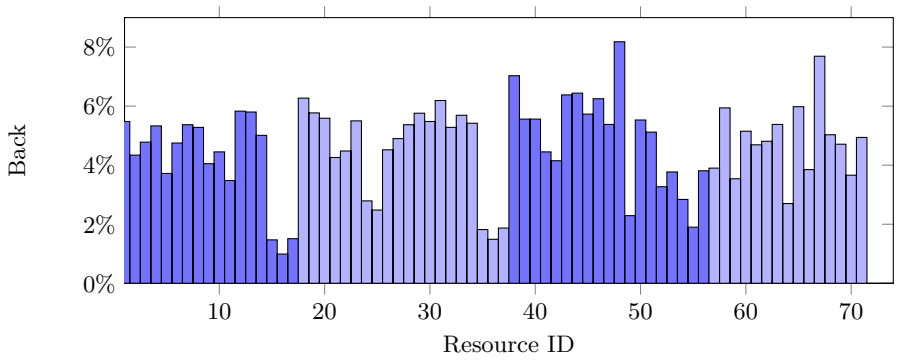

Fig. 17. 'Back' for each resource in the Nano course

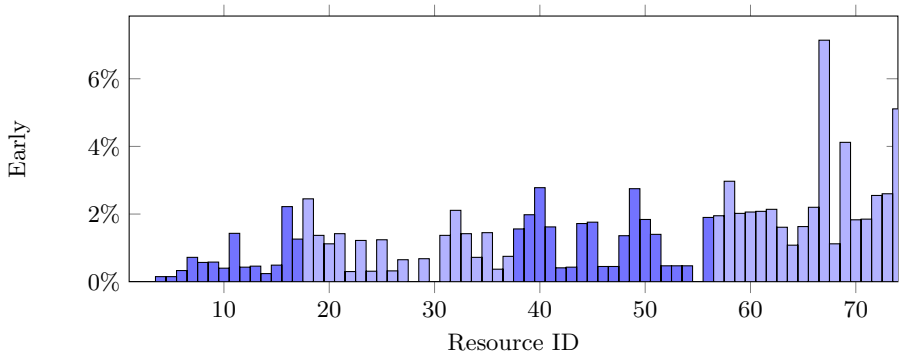

Fig. 18. 'Early' for each resource in the Nano course

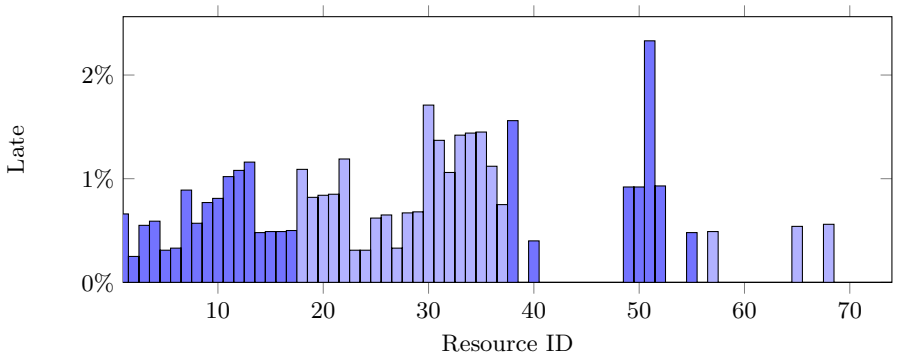

Fig. 19. 'Late' for each resource in the Nano course

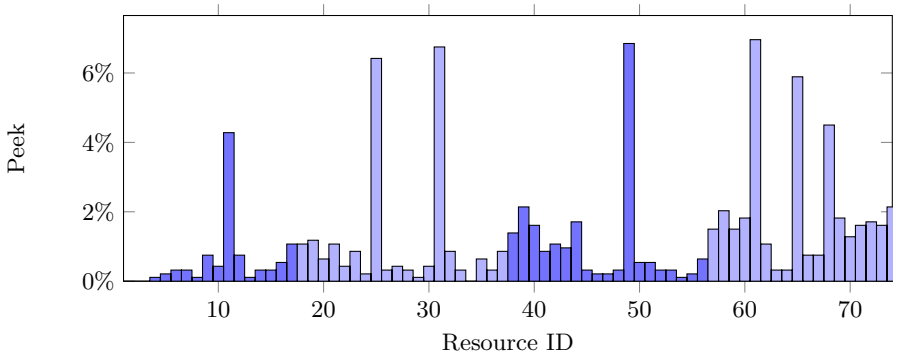

Fig. 20. 'Peek' for each resource in the Nano course

A.3 Features for the Ultra course

Figure 21 contains the graph of the number of learners that are still active at each resource.

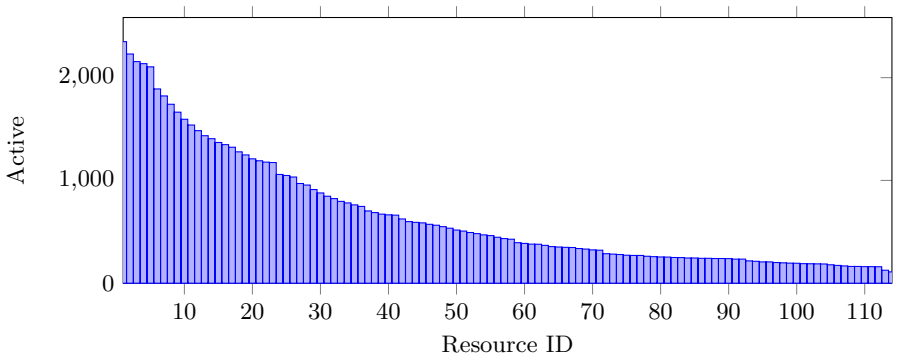

Fig. 21. Active learners for each resource in the Ultra course

Figures 22, 23, 24, 25, and 26 contain respectively the graphs of the features 'drop', 'skip', 'back', 'early', and 'late' for each resource. They are features computed as a proportion of the number of active learners. Figure 27 contains the graph of the feature 'peek', which is computed as a proportion of all learners. In these six figures, the change of week is shown by changing the shade of blue. The even weeks are lighter than the odd weeks.

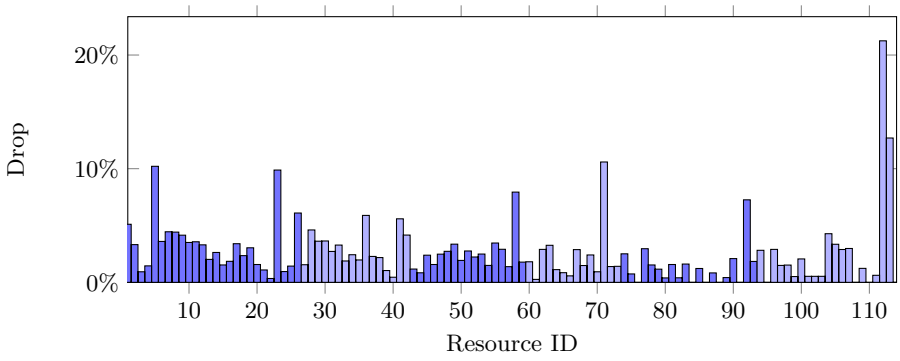

Fig. 22. 'Drop' for each resource in the Ultra course

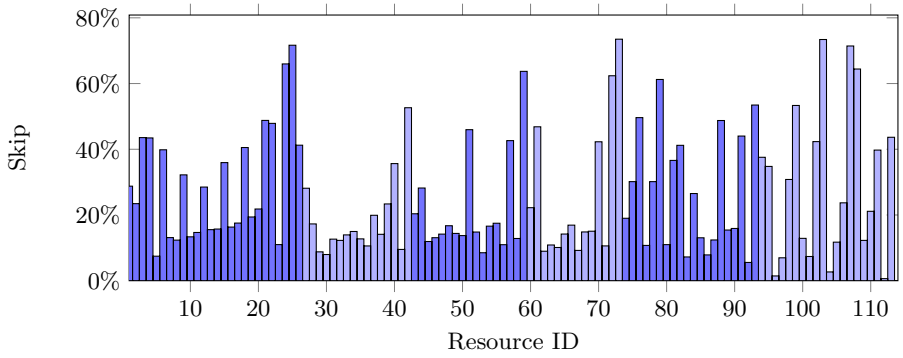

Fig. 23. 'Skip' for each resource in the Ultra course

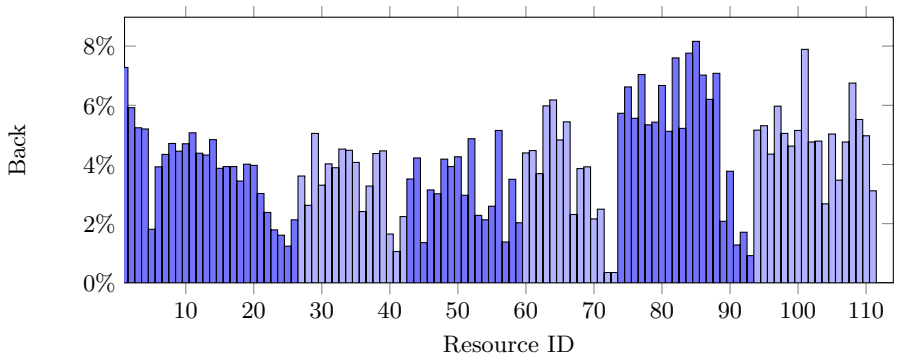

Fig. 24. 'Back' for each resource in the Ultra course

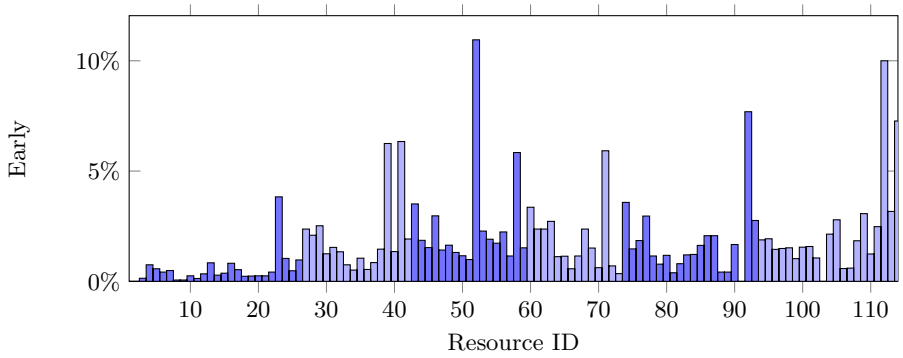

Fig. 25. 'Early' for each resource in the Ultra course

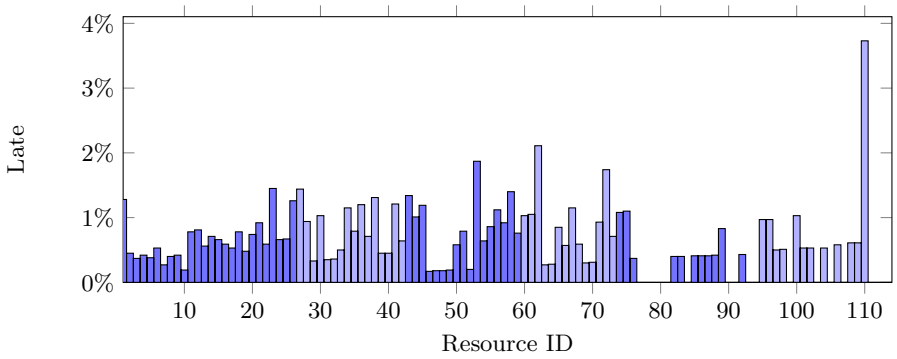

Fig. 26. 'Late' for each resource in the Ultra course

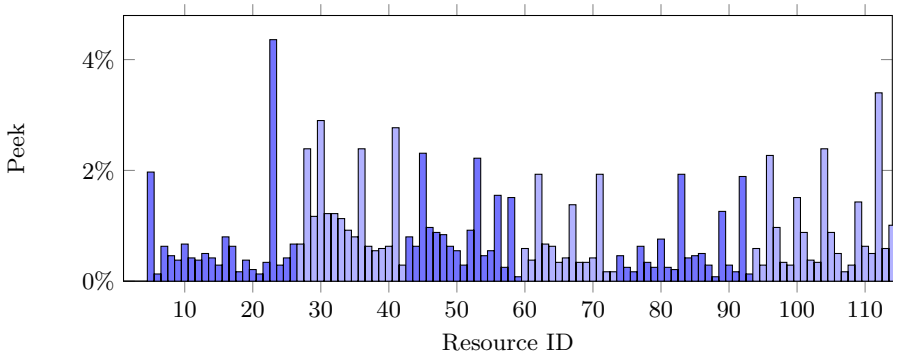

Fig. 27. 'Peek' for each resource in the Ultra course

A.4 Features for the Converted course

Figure 28 contains the graph of the number of learners that are still active at each resource.

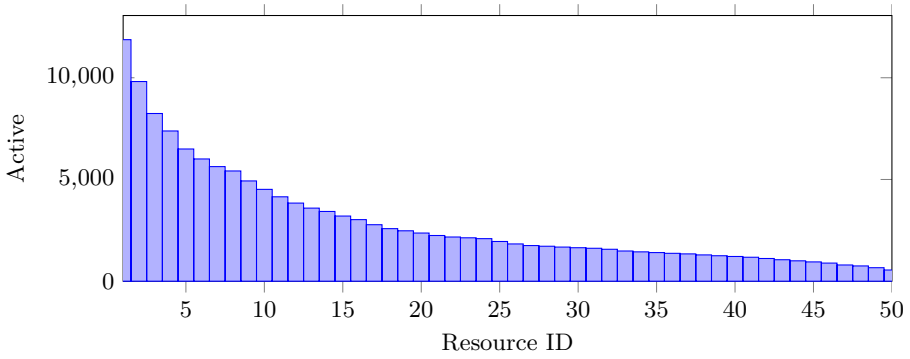

Fig. 28. Active learners for each resource in the Converted course

Figures 29, 30, 31, 32, and 33 contain respectively the graphs of the features 'drop', 'skip', 'back', 'early', and 'late' for each resource. They are features computed as a proportion of the number of active learners. Figure 34 contains the graph of the feature 'peek', which is computed as a proportion of all learners. In these six figures, the change of week is shown by changing the shade of blue. The even weeks are lighter than the odd weeks.

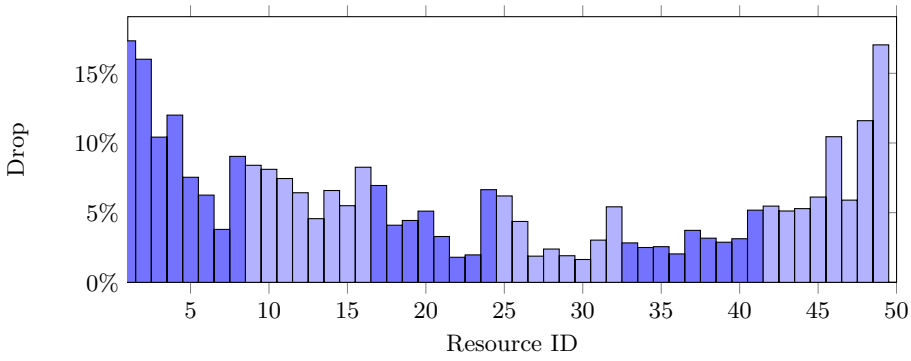

Fig. 29. 'Drop' for each resource in the Converted course

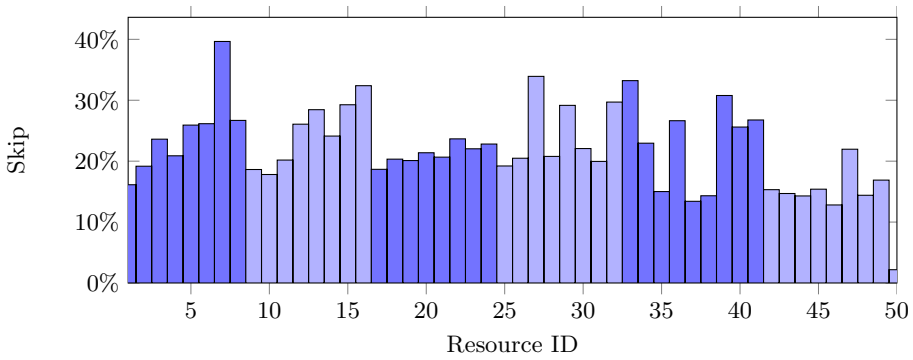

Fig. 30. 'Skip' for each resource in the Converted course

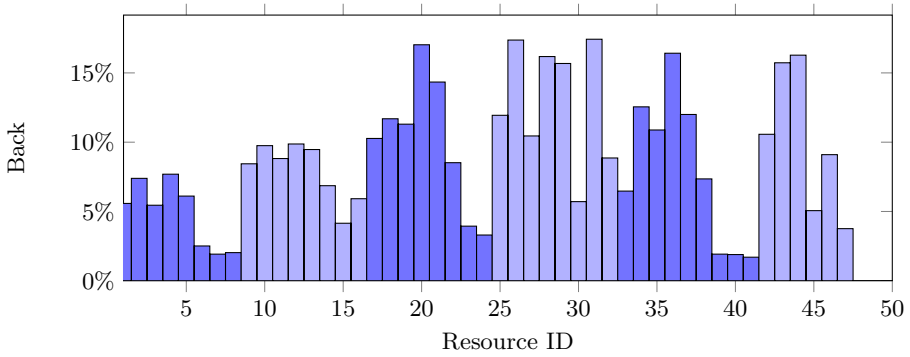

Fig. 31. 'Back' for each resource in the Converted course

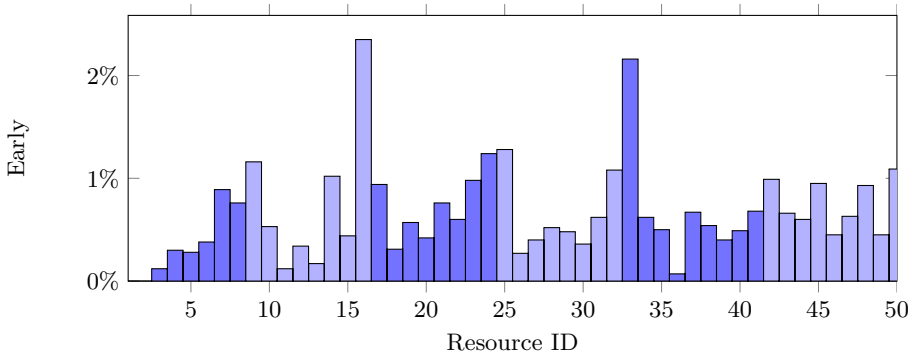

Fig. 32. 'Early' for each resource in the Converted course

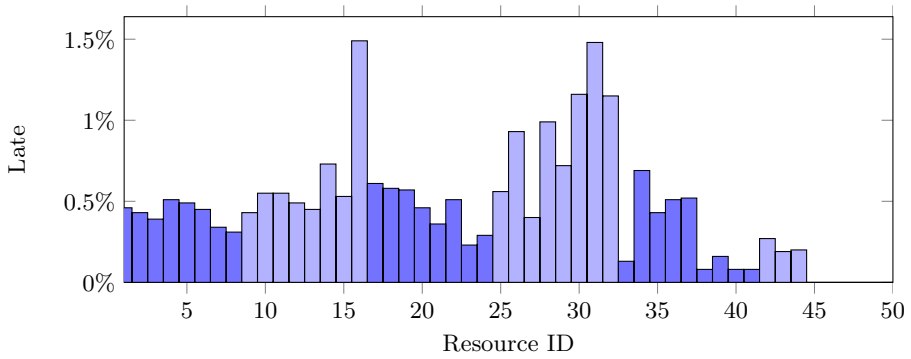

Fig. 33. 'Late' for each resource in the Converted course

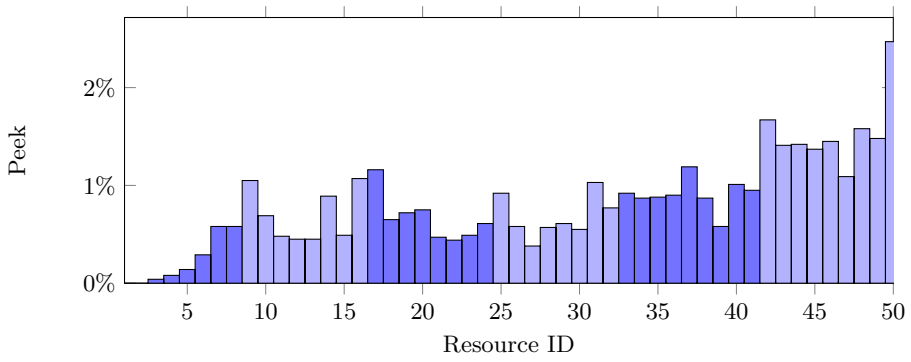

Fig. 34. 'Peek' for each resource in the Converted course

A.5 Features computed with alignments - ProM course

Figures 35, 36, and 37 respectively present features similar to the 'drop', 'early' and 'late' features, when obtained through the method of alignments. They are obtained from the same data as the features presented in A.1.

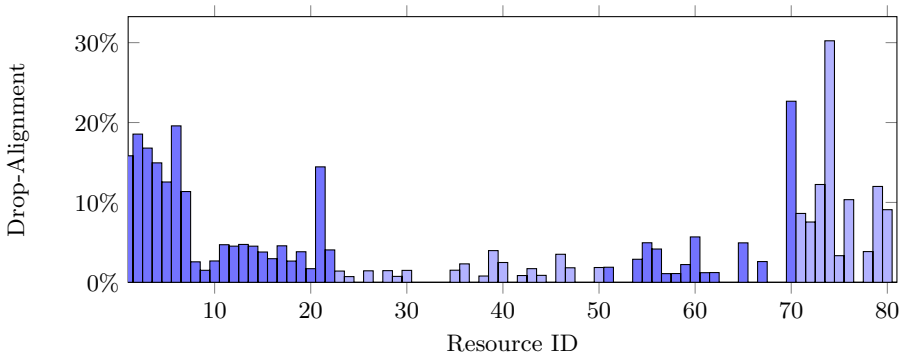

Fig. 35. Alignment-based 'drop' for each resource in the ProM course

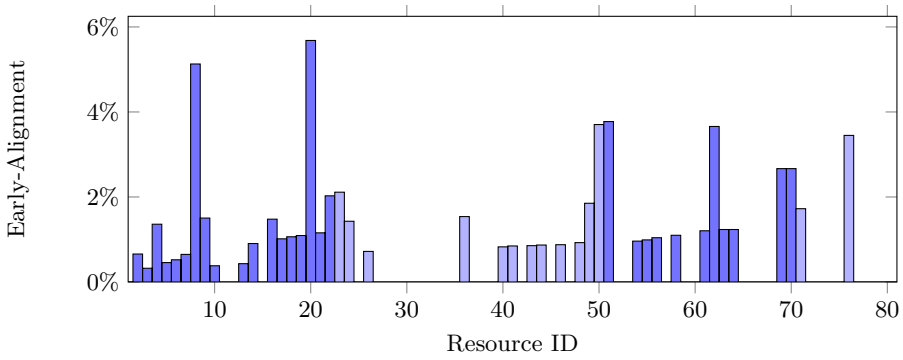

Fig. 36. Alignment-based 'early' for each resource in the ProM course

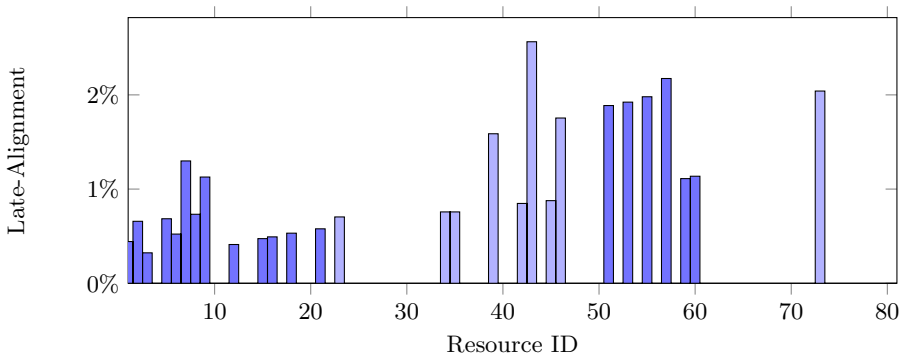

Fig. 37. Alignment-based 'late' for each resource in the ProM course

A.6 Features computed with alignments - Converted course

Figures 38, 39, and 40 respectively present features similar to the 'drop', 'early' and 'late' features, when obtained through the method of alignments. They are obtained from the same data as the features presented in A.4.

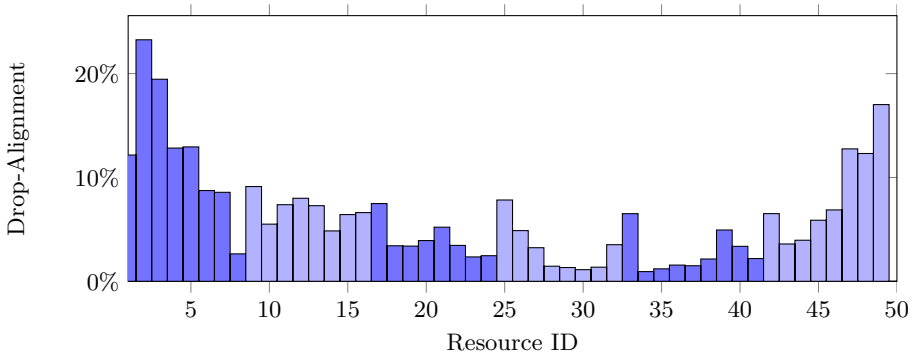

Fig. 38. Alignment-based 'drop' for each resource in the Converted course

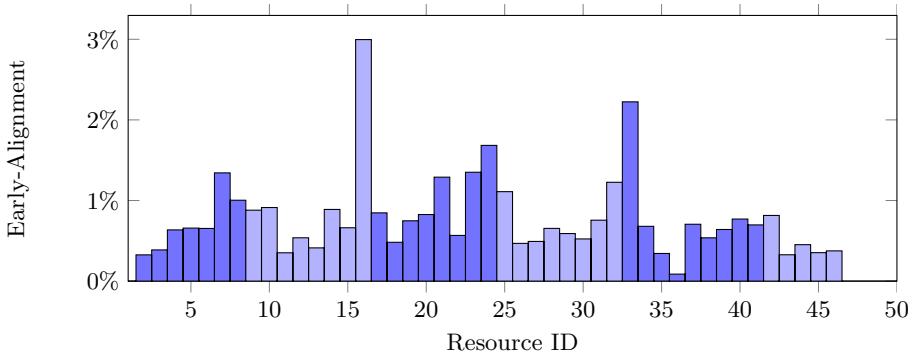

Fig. 39. Alignment-based 'early' for each resource in the Converted course

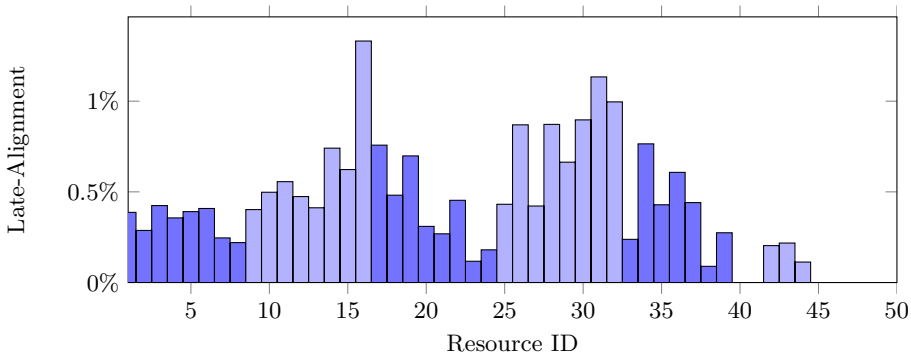

Fig. 40. Alignment-based 'late' for each resource in the Converted course