Exploring engagement profiling in MOOCs through Learning Analytics: The Open edX Case

Author: Andrey Belinskiy

Supervisors: Prof. Barbara Wasson and Dr. Mohammad Khalil

Department of Information Science and Media Studies
University of Bergen
2021

Acknowledgments

First of all, I want to thank my friends, relatives and fellow students who always supported me during these daring times. Secondly, I thank my supervisors from the SLATE centre, Dr. Mohammad Khalil and Prof. Barbara Wasson, for helping me in carrying out the research and guiding me along the way, and Gleb Belokrys for the provided technical support. Lastly, I am very grateful to the University of Bergen and the Faculty of Social Sciences for providing the opportunity to work together with and learn from many interesting and truly knowledgeable people during my studies.

Abstract

The enormous amount of data being generated daily, requires effective and efficient ways of processing and analysing in order to extract useful information and form meaningful conclusions. Learning Analytics is a set of methodologies and practices that uncover such information from educational data. The research in this thesis explores the addition of a Learning Analytics feature to the context of a Learning Analytics tool that aids instructors using the online Massive Open Online Course (MOOC) platform, Open edX. This is done through the development and evaluation of a working artefact that supports profiling of students according to their activity throughout the course, alongside the visualizations, which represent said activity. As a result, the thoroughly demonstrated process of the artefact creation and feedback collection from the instructors shows the potential of Learning Analytics methods when applied to Open edX tracking data. Several practical features for creating different engagement groups, together with the visualizations, are conceptualized, implemented and evaluated, and are positively assessed by the target group of instructors. In addition, the challenges that were encountered in the period of the development, are presented, together with the suggestions to overcome them. Finally, a few extra features are outlined for future work, which could expand the existing functionality even more and bring additional knowledge to this research area.

Table of Contents

Table of Figures

Table of Tables

Chapter 1

Introduction

The amount of data and information is growing rapidly each day. Therefore, it is very important to keep up with its volume and variety, and to find ways of processing it in a scalable and efficient manner. Otherwise, we may miss a number of crucial observations and insights, which could be used to enhance and increase the effectiveness of the existing practices, or even completely new and innovative ways of solving existing and future problems. Thus, the research, such as this thesis, which is aimed at finding new and efficient methods of working with large amounts of student data, and producing meaningful results is crucial.

During past years online learning and digital platforms for online education have been steadily gaining more and more popularity (Bozkurt et al., 2016). The same trend applies to higher education organizations like universities and schools in the form of MOOCs and Learning Management Systems that are being integrated into the core education process. This in-turn brings both positive opportunities and new challenges for the instructors (Len-Urritia et al., 2018).

Another result of the usage of digital learning tools is the vastly increased amount of generated data. This data includes a magnitude of different types of observations and statistics about many distinct actions and events that happen during the period of education. To be able to extract meaningful information from this data a number of different data mining and data-processing methods can be used. These methods are a part of the Learning Analytics (LA) field, which mainly focuses on collecting and analysing the data about learners and their learning environment in order to understand and improve the educational process (Khalil & Ebner, 2015).

One of the common methods in LA is clustering (Aldowah et al., 2019). It consists of grouping the students in the course based on several different events, which are recorded in the system when students interact with the course. Such events include, among other things, video interactions, forum activity and solving tasks. As a result, the behavioural patterns of each student can be identified, which provide tailored feedback to the instructors and allow them to assess the current state of the course as a whole, as well as to intervene to support struggling students and encourage active learners. Clustering, however, requires relatively large datasets in order to produce meaningful results, and might not be suitable for all courses implemented as MOOCs. This is especially crucial in the case of Small Online Private Courses (SPOCs), which usually have a relatively small number of students enrolled in them. To solve this issue, that is the lack of tracking data for clustering in the case of a SPOC, an alternative solution is proposed in this research. This solution consists of forming student engagement profiles, which capture the activity of the course participants and allow the instructors to analyse this information and make appropriate course-related decisions. These profiles can be dynamically customized through changing the weights of course activities, depending on which activity is considered more or less important by the instructor.

Thus, this research aims to support instructors in understanding student behaviour and engagement in the Open edX MOOCs. By using LA to profile students with similar behaviours the instructor receives useful information about their students. This information is presented in the LA tool OXALIC (Khalil & Belokrys, 2020), which has been developed for use with Open edX MOOCs. In order to meet this goal, the research will 1) determine how to develop student profiles based on student activity data in Open edX MOOCs, 2) develop an artefact to be integrated in OXALIC, 3) investigate how to enable an instructor to manipulate the weighting of variables used in profiling, and 4) explore how these profiles can be used by instructors to make course related decisions.

1.1 Research questions

In order to reach the aim of the research, the following research questions were formulated for the thesis.

RQ1. How to identify engagement in Open edX MOOCs?

RQ2. What student profiles emerge through LA when it is applied to the activity data, and how can this be presented to instructors in the LA tool OXALIC?

RQ3. How do instructors use these student profiles to make course-related decisions?

1.2 Thesis outline

The outline of this research project is presented below:

Chapter 2: Literature review provides an overview over what has been found in the literature regarding LA and clustering.

Chapter 3: Open edX platform contains a description of the online course platform and its tracking logs, which are used for the main part of the project, alongside with the short description of the existing LA tool, OXALIC, which uses the aforementioned platform.

Chapter 4: Methodology and methods describes the main system of methods that were used for this study.

Chapter 5: LA algorithms presents a detailed overview of the proposed algorithm for profiling based on the Open edX tracking data. It also summarizes the details of the visualization concept for the profiling algorithm.

Chapter 6: Artefact development describes how the artefact was developed, which tools and technologies were used and how the data was processed. It also
overviews the process of development of the visualization part for the proposed algorithm, alongside the tools and technologies, which were used for that.

Chapter 7: Evaluation summarizes the results of the conducted evaluation, including the details about how it was performed, and the feedback from the users regarding the usability and performance of developed artefact.

Chapter 8: Discussion contains the general overview and thoughts of the research, as well as the answers to research questions.

Chapter 9: Conclusion and future work is a brief summary of what has been achieved by this research, and several thoughts and ideas regarding the future work on this subject.

Summary

This chapter provided an introduction to the thesis with general information about the goal of the research, as well as the overview of the chapters of the thesis. It allows the reader to have a quick grasp over the whole thesis and understand its structure.

Chapter 2

Literature review

To outline the context for the research, the current state of LA in higher education and its application in different educational scenarios has been explored. To achieve this, resources such as The Web of Science and Google Scholar have been used to find the relevant scientific articles. A general overview about the use of LA for education is given and an understanding of where more research is needed is identified. Finally, insights about the application of LA and Educational Data Mining (EDM) in MOOCs are gained. The review of the literature relevant to this research is presented below.

2.1 Review methodology

The main method that was applied when conducting the literature review is desk research, which is also known as secondary research. The goal of desk research is to find the already existing information about the main area a researcher is focusing on in their studies. This is a necessary step that helps researchers to understand the current state of the study area and to utilize this information to supplement and support their own endeavours.

In order to find the existing information and knowledge about the topic of this research in the form of scientific articles, the “Web of Science” and “Google Scholar” search tools were mainly used. The process of finding and identifying the literature relevant to the research is shown in Figure 1, which depicts the PRISMA flow diagram.

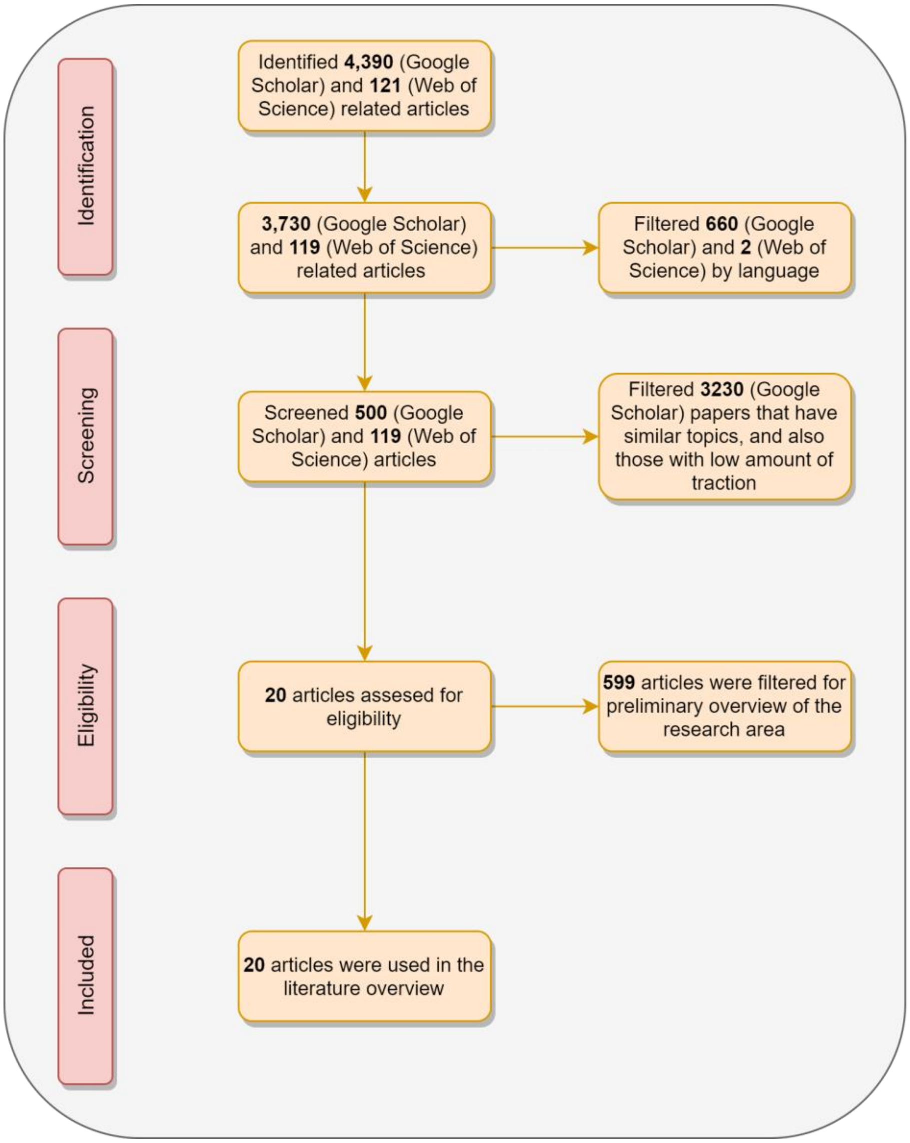

Figure 1. PRISMA flow diagram.

The steps that were taken in order to find the articles are as follows. First, a batch of articles was identified using the combination of keywords and logical operators. The query that was used for the first iteration consists of the following elements - *("Learning Analytic*" OR "EDM" OR "Educational Data Mining") AND ("MOOC" OR "Massive Open Online Course*" OR "Small Private Online Course*" OR "SPOC")*. The inclusion of SPOCs as well as MOOCs into the search query is due to the fact that this is another popular type of online courses used in higher education and therefore it is relevant to the research topic. Next, some papers were excluded since they were not in English language. After that, the number of papers was filtered due to having similar topics or low traction, that is, low number of citations, taking into account that several years have passed since their publication. This may be considered as a questionable metric of relevance for the papers, but in this case the papers that were left after this step were enough for getting meaningful information about the research area. Next, these papers were screened and looked through to identify their relevance. During this step, some papers were excluded because they focused on very specific aspects and techniques that were not as relevant for this study. As a result, a relatively small number of papers was left for the literature review.

2.2 Learning Analytics in higher education

The first paper named “The current landscape of learning analytics in higher education” (Viberg et al., 2018) presents the results of a literature review of 252 papers. The main research question that the authors try to answer in their study is “What is the current scientific knowledge about the application of learning analytics in higher education?” (Viberg et al., 2018, p. 99). The study covers the papers that were published in the period from 2012 to 2017 as well as the proceedings from the “Learning Analytics & Knowledge” conference starting from 2018. This conference is considered a “premier research forum in the field, providing common ground for all stakeholders in the design of analytics systems to debate the state of the art at the intersection of Learning and Analytics — including researchers, educators, instructional designers, data scientists, software developers, institutional leaders and governmental policy makers”.[1]

The paper provides several key insights which can be helpful for understanding the overall picture of the LA application in higher education:

- LA can be considered a maturing field, based on the fact that 26% of the papers that were included in the review are categorized as “theory use” rather than “theory generating” studies (Viberg et al., 2018).
- The general focus of research in LA is shifting from predictive methods to finding the relationships between different components and agents in higher education, and to collect, formalize and visualize the data for humans to make decisions based on the processed data (Viberg et al., 2018).
- The potential of LA application for enhancing the results and experience of education is high, but “there is little evidence (9%) that the research findings demonstrate improvements in learning outcomes” (Viberg et al., 2018, p. 108). The authors therefore stress out that it is crucial to understand how to transfer this potential into actual results.

The second paper named “Educational data mining and learning analytics for 21st century higher education: A review and synthesis” (Aldowah et al., 2019) surveyed 402 articles about EDM and LA. The authors acknowledge the potential of EDM and LA in higher education and aim to present a thorough review of different techniques and methods that are used in this field.
Here are several main excerpts extracted from the paper:

- Classification and clustering are the most commonly applied data mining techniques in higher education (Aldowah et al., 2019). Classification can be described as a technique that assigns collected data to one or several classes. It can be used, for example, to

[1] The Society for Learning Analytics Research (SoLAR). *International Conference on Learning Analytics & Knowledge (LAK)*. Retrieved from https://www.solaresearch.org/events/lak/.

predict a certain outcome of student activity, or to understand the overall behaviour of the students based on their activity in the system. Clustering is a method of grouping the classes that have some similarity into bigger entities that are called “clusters”. This way, for example, the student that belongs to a certain cluster can be given a similar activity as other students in the same cluster.

- “The applications of EDM/LA are a growing phenomenon of the 21st century higher education” (Aldowah et al., 2019, p. 29). It is stated that the amount of research and number of studies were progressively increasing over the period from 2014 to 2019. Therefore, it can be concluded that this field is actively developing, and different opportunities should be available for those who aim to advance the current state-of-the-art in this research area.
- The authors conclude that “the application of EDM/LA can provide significant benefits, and therefore [the authors] urge higher education institutions to adopt them where feasible” (Aldowah et al., 2019, p. 31). Indeed, based on the presented data it can be stated that the usage of these techniques has big potential for enhancing learning outcomes for students as well as providing better overview of student activity for course creators by utilizing various advanced visualization tools.

The third article “Features students really expect from learning analytics” (Schumacher & Ifenthaler, 2018) is a qualitative study that involved 20 university students, and a total of 216 students for quantitative study to supplement the results of the first qualitative part. The goal of the study is to understand the expectations of students to accept different LA practices and techniques during their education period as well as the willingness to use them. The paper also emphasises the importance of self-regulated learning “as a vital factor for learning success” (Schumacher & Ifenthaler, 2018, p. 398). In this regard, different LA techniques can be beneficial due to the ability to process various data and, as a result, provide students with useful and meaningful feedback about their learning progress.

The main findings are the following:

- The students generally seem to have a positive reception of LA applications during their studies. However, the students prefer to avoid comparisons of their results with other learners (Schumacher & Ifenthaler, 2018).
- The paper suggests focusing more on LA features that support self-regulated learning when designing the learning environment for students. This is supported by the fact that three out of five features that students were willing to accept are “repetition of learning content, prompts for selfassessment, and further learning recommendations to complete a course” (Schumacher & Ifenthaler, 2018, p. 405).

Based on the brief overview of these three papers it can be concluded that LA in higher education has great potential and can be very beneficial for researchers, instructors and students. It is also mentioned that MOOCs alongside other online learning environments are one of the reasons for the progressively increasing amount of interest and research in the field of LA (Aldowah et al., 2019). Therefore, in the following part of this overview, this approach to online learning will be explored more thoroughly. EDM will also be a focus of literature analysis, since its methods and techniques can present meaningful results in the context of processing the data collected in an online learning environment (Aldowah et al., 2019).

2.3 Learning Analytics in massive open online courses

In order to understand how LA methods are applied in MOOCs, and what information is generated in the process, a number of articles was selected for the review. Additionally, a set of guiding questions was formulated to aid with the direction of the review, and to further explore the RQ1, which was mentioned in Chapter 1, in more detail. These questions are the following:

- How can Learning Analytics techniques be used to explore data from MOOCs?
- What can Learning Analytics reveal out of the raw level data of MOOCs?
- Can Learning Analytics be used to support MOOC's stakeholders in decision-making? How is that possible?

In this part each of these directions will be explored and the findings in the literature regarding these questions will be presented. It can also be said that these questions are closely connected and overlap with each other, and the literature often covers more than one question. Therefore, the categorization of papers between them in the next part is arbitrary and not in any case absolute.

2.3.1 How can Learning Analytics techniques be used to explore data from MOOCs?

This research question was partly covered previously in this review. We learned that there are many different LA methods and techniques that can be used for getting useful information from educational data. Now the goal is to look closer at the specific type of educational platform, namely MOOCs, and explore the applications of LA in this area.

As it was mentioned before, one of the major techniques that is used for processing the educational data is clustering. The application of clustering in MOOCs is therefore a viable approach to extract meaningful information from the data that is generated during the MOOC.

The authors of the first paper "Clustering patterns of engagement in Massive Open Online Courses (MOOCs): the use of learning analytics to reveal student categories" (Khalil & Ebner, 2017) focus on the task of grouping students based on their engagement. As a result, the authors present four groups of students that they were able to identify: "Social", "Gaming the System", "Dropout" and "Perfect Students", based on the activity and level of engagement in the course. One of the benefits of such grouping is that the instructors then can make an intervention and influence the whole group of students into changing their behaviour (Khalil & Ebner, 2017). This alone provides a great benefit for instructors and can change the way they interact with their students.

In the second paper "Analysing Structured Learning Behaviour in Massive Open Online Courses (MOOCs): An Approach Based on Process Mining and Clustering" (van den Beemt et al., 2018) the authors demonstrate another application of LA in MOOCs, namely they try to discover the correlation between the distribution of students' weekly activities and their success or failure in the course. They found out, for example, that there is no confirmation that switching between assignments from different parts of the course in contrast to following assignments in success, leads to better learning outcomes (van den Beemt et al., 2018). It is also mentioned that the results of LA can be very beneficial for instructors for understanding the behaviour of their students.

2.3.2 What can Learning Analytics reveal out of the raw level data of MOOCs?

This part focuses on the so-called "raw" level data that is generated by MOOCs. This includes, for example, the number of clicks the student makes on a certain page of the course, or number of pauses at certain points in time of the video playback. This data is usually presented in a more technical way and does not immediately provide insights about the student's activity. Therefore, several applications of LA, which help to transform this type of "raw" data into something that can be used for making decisions or providing certain statistics, will be explored.

The authors of the paper "Using learning analytics to evaluate a video-based lecture series" (Lau et al., 2018) leverage the technical data about video lectures in a medical course to then process it and understand what valuable information can be extracted. One of the conclusions that was derived is that learners in the medical field may prefer longer videos with more details than ordinary students (Lau et al., 2018). The authors also propose a model for evaluating the video-based lectures part of the MOOC and possible solutions to how to increase the retention of the audience.

In the next paper "Mining MOOC Clickstreams: Video-Watching Behavior vs. In-Video Quiz Performance" (Brinton et al., 2016) the authors propose two frameworks based on mathematical models that use students' raw data in a form of clickstreams, and then analyse the results. The authors observed, among other things, similarities in students' behaviour when interacting with videos and the correlation between these behavioural patterns and success or failure in quizzes (Brinton et al., 2016).

Two more papers related to usage of "raw" data in LA and its implementation in a form of working tool were discovered, namely "Scaling to Massiveness With ANALYSE: A Learning Analytics Tool for Open edX" (Ruipérez-Valiente et al., 2016) and "edX Log Data Analysis Made Easy" (Torre et al., 2020). These papers describe two different tools that use LA methods and techniques to provide main stakeholders with useful insights. They will be covered later.

2.3.3 Can Learning Analytics be used to support MOOCs' stakeholders in their decision-making? How is that possible?

This question was partly covered in (Schumacher & Ifenthaler, 2018) in that students can use LA to organize and plan their work while reflecting onto their current progress with the course. Another application of LA that may benefit both students and instructors is presented in the paper "Sociograms: An Effective Tool For Decision Making in Social Learning" (Zorrilla & de Lima Silva, 2019). The authors show how instructors and students can change their behaviour based on visualizations of their social interactions, which are presented in a form of graphs. For example, "teachers could try to activate discussion about topics with a few or no messages or redirect their target; and students could check activities where they have not yet participated" (Zorrilla & de Lima Silva, 2019, p. 670). Since MOOCs often include different social elements like forums and chats as a part of the course, the addition of this type of visualizations for both student and instructor dashboards can be very useful in providing the analysis and measurement of the relationships between the participants and the decisions that are made based on this information (Zorrilla & de Lima Silva, 2019).

Finally, the authors of the paper "Using Learning Analytics to Improve MOOC Instructional Design" (Shukor & Abdullah, 2019) propose key aspects that course creators and instructors should keep in mind when designing the MOOC. The results are based on the data collected from two courses that were publicly available for learners. Some of the key concepts that were found are the importance of a useful and functional home page for the course, and the relatively big role of self-reflection and the ability to evaluate the course (Shukor & Abdullah, 2019).

2.4 edX, Open edX and existing tools

The next step after establishing the preliminary research questions is to look for actual implementations of the LA methods and techniques in the case of edX and Open edX. The goal of this step is to understand the capabilities of these implementations, and to what extent they utilize LA to provide the stakeholders with the necessary information.

First of all, in order to generate the data for LA to process there should exist some sort of a platform that will provide the necessary tools and technologies. One of such platforms is Open edX, a non-profit open-source ecosystem that allows practically anyone to use it as a set of instruments for creating and managing different types of online courses, including MOOCs. In 2018 there were more than 1500 websites that used this platform, with more than 18000 courses total (DjangoCon US, 2018). edX, on the other hand, is a commercial version of said platform that has the similar architecture but has a number of additional features that are available for customers, such as for example different, more convenient structure of data logs generated by the platform. The Open edX platform is overviewed in more detail in the following chapters.

To effectively and efficiently process the data that is generated by the Open edX ecosystem, several tools were developed and implemented. Two such tools described in (Ruipérez-Valiente et al., 2016) and (Torre et al., 2020) will be overviewed below.

In “Scaling to Massiveness With ANALYSE: A Learning Analytics Tool for Open edX” (Ruipérez-Valiente et al., 2016) the authors propose a tool that focuses on presenting the visualization for the main stakeholders, i.e., students and instructors. The architecture of the tool is presented in Figure 2.

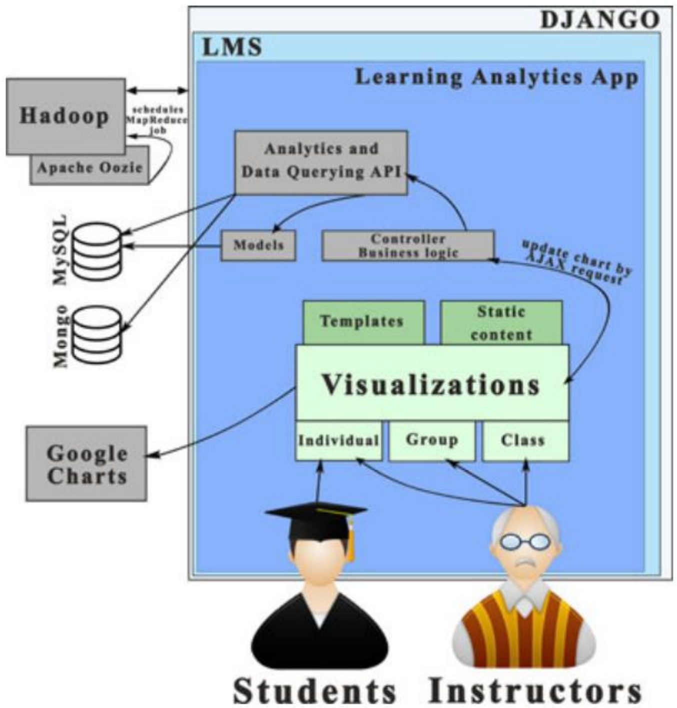

Figure 2. Architecture of the ANALYSE tool (Ruipérez-Valiente el al, 2016).

Some of the key insights that the authors provide are:

- The biggest challenge that the authors had to deal with is the massiveness of the data that is generated by the Open edX platform. This includes data such as interaction with different elements on the course page, interactions with video content, etc. The authors were able to solve it with the use of specific technologies, namely “MapReduce”. The main principle of this technology is to divide the data into smaller pieces and then process them in parallel using a cluster of devices. The results of this processing are then combined into one entity which represents the result of processing the whole initial piece of data. It is usually a good choice for dealing with large amounts of data in a timely manner (Ruipérez-Valiente et al., 2016).
- The tool was able to produce a number of rich visualizations using the processed data. One of the most interesting and insightful of them is the one that shows statistics about video interaction events. For example, it is able to show which part of the video was

replayed more than others, or how much time was spent on videos in comparison with the assignments or quizzes.

The authors of “edX Log Data Analysis Made Easy” (Torre et al., 2020) also present the tool called “edX Log file Analysis Tool” (ELAT) for the same purpose of processing the edX data, and it is a very recent one. The architecture of the tool is presented in Figure 3.

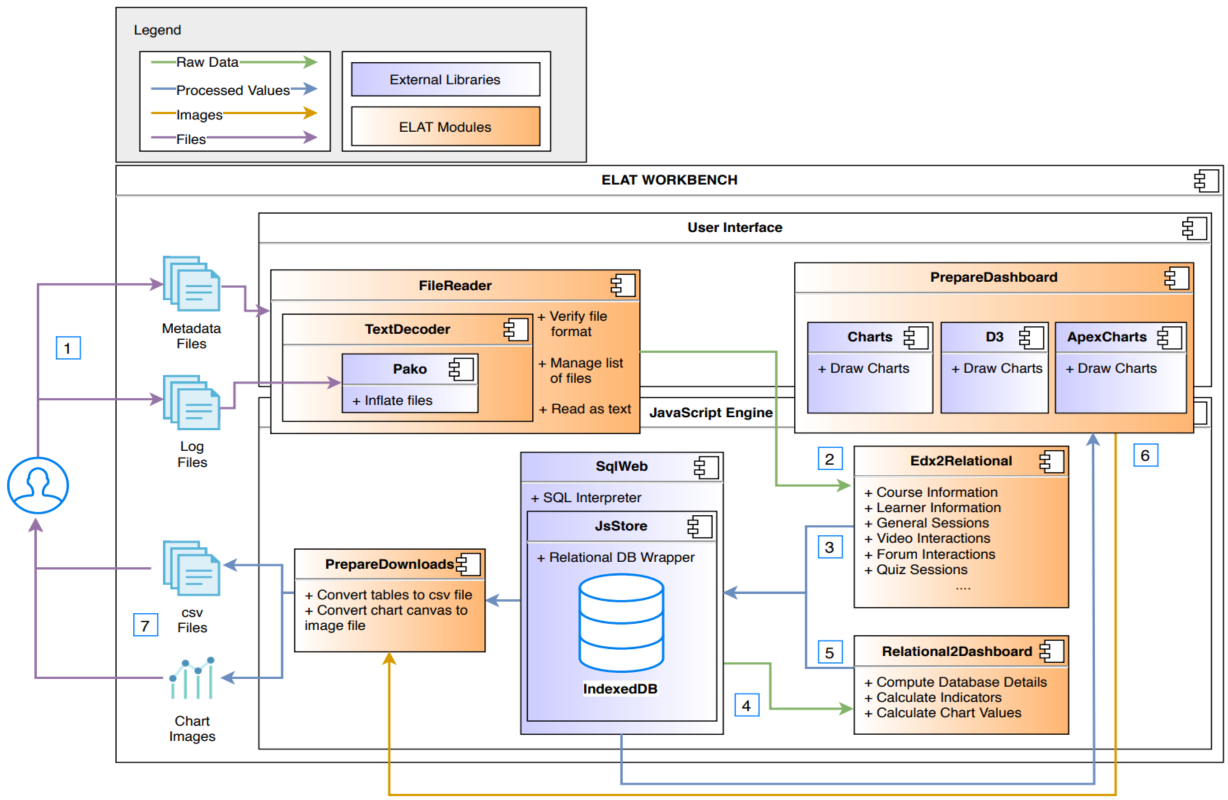

Figure 3. Architecture of the ELAT tool (Torre et al., 2020).

The main insights that can be gained from this paper are:

- The authors make an overview of the existing tools and emphasize that most of them require technical knowledge and significant time for setting up the system. The tool they propose is user-friendly and can be run right from the go in an internet browser. This way instructors and researchers can get relatively quick access to the visualizations of the data and analyse them.
- In addition to generating visualizations, the tool can build semantic entities. For example, by processing the data logs the tool can generate a “study session” that represents several chained events the student participated in during a certain period of time. This in-turn provides a possibility for deeper analysis of students’ behaviour and learning patterns, and presents additional triggers for interventions (Torre et al., 2020).
- The tool was evaluated in a form of analysing how it works with large amounts of data, as well as by seven learner research experts. In both cases the results were positive: (i) the

tool processed the data in a reasonable amount of time and (ii) the experts acknowledged the usefulness of the information that was generated by the tool.

After overviewing these two papers it can be concluded that (i) the task of applying LA for extracting meaningful information from “raw” data is quite relevant, and that (ii) different tools that focus on this area continue to emerge as the time goes by. It is also important to mention that both tools are available as open-source projects, which allows any interested party to use parts of the tools or even the whole system and apply them in their particular case.

2.5 Research context

Based on the conducted literature review, an approximate research context has been established. The potential that emerges from the application of LA in education is undoubtful. However, there is not enough factual evidence that the generated information does indeed directly lead to better results in the process of education (Viberg et al., 2018). This research area will therefore be explored and observations and conclusions, which will be drawn during this master thesis, will be contributed to the knowledge base.

Another research area is the vast amount of available LA techniques that are tailored to solve very specific tasks. It is not very clear which method is more effective than another, and for which situations it is best suited for. One of the areas of LA techniques will be focused on and examined in this thesis in order to analyse how these techniques are used and what are the nuances, which emerge in different use-cases of said techniques.

2.6 Clustering based on engagement

After conducting an overview over the existing state of the research area, one specific part of it was chosen for a deeper overview, namely clustering of students based on their engagement, since it is one of the most common data mining techniques applied in educational environments (Aldowah et al., 2019). Thus, it can be assumed that there is enough existing knowledge and information available for analysis, and for identifying the areas, which require more research. Additionally, by focusing on one problem area, a more specific research context can be outlined. For this purpose, five additional papers were found and overviewed, with the focus on the size of the datasets, the resulting clusters of students, as well as the exact methodologies used to group students depending on their engagement. One of the main purposes of this overview is to assess the viability of applying the advanced EDM techniques, such as machine learning, and clustering more specifically, in order to get a meaningful result considering the small sample size of the data available, due to the relatively small number of students present in each separate university online course. The review of these papers is presented below.

In the paper “Research on Clustering Mining and Feature Analysis of Online Learning Behavioral Data Based on SPOC” (Zhang et al., 2018) the authors utilize the dataset formed during a SPOC with 700 learners. The main machine learning algorithm that the authors utilize to form the clusters is k-means, which seems to be the most commonly used one throughout the other works in this particular area, as it will be presented further. To perform clustering analysis to identify learners of different styles, the authors select four indicators: (i) number of posts and replies, (ii) final scores, (iii) total duration of watched videos, (v) the number of videos viewed. The resulting clusters are presented in Table 1.

Table 1. The resulting clusters in (Zhang et al., 2018).

Cluster	Description
weak-cognitive learners (NO.0)	Those with high video viewing rates, long duration but low final scores.
self-conscious learners (NO.1)	The excellent learners who have completed the indicators that do not count toward achievement.
short-cut learners (NO.2)	Those with a higher final score, but who have a low completion rate of indicators that do not count towards achievement.
lazy learners (NO.3)	The learners who do not have high-scored indicators.

The authors also use a hierarchical clustering merging algorithm that determines each sample point’s similarity by calculating the distance between each category of data points and all data points. The smaller distance is, the higher similarity will be. Additionally, the authors adopt Ward, that is the square sum of deviations, as the main method to measure the distance between two clusters.

In “Moving Through MOOCS: Pedagogy, Learning Design and Patterns of Engagement” (Ferguson et al., 2015) the authors use The Open University (OU) study presented in (Ferguson & Clow, 2015) as the foundation for their research project, with four datasets available from this study. The original method in that paper focuses on engagement with content and assessment, and results in the following four groups: (i) “on track” if students submitted assessment in the week it was set, (ii) “behind” if students completed an assessment after the week in which it was set, (iii) “auditing” if students engaged with content but not with the assessment, (v) “out” if students did not participate in a course week. This method, however, did not work well for the data that the authors had access to, since the courses were based on a different learning platform, “Future Learn”. The data contained a lot of social interaction elements at each step of the courses, like forum interactions and discussions. The authors therefore came up with a modified method - create engagement profiles for learners that reflected engagement with content, with assessment and with discussion. The method itself will be briefly described shortly after. The resulting profiles are presented below:

- **Samplers** visited a course briefly.
- **Strong Starters** left after the first week's assessment.
- **Returners** completed assessments in the first two weeks, then left.
- **Mid-way Dropouts** completed 3–4 assessments before leaving.
- **The Nearly There** cluster completed most assessments but left early.
- **Late Completers** completed most assessments but were either late in submitting these or missed some.
- **Keen Completers** engaged actively throughout.

MOOCs in the OU study were mainly eight-week courses with an assessment point at or near the end of each week. Second MOOC ran for a shorter period, and here the "Mid-way Dropouts" cluster was replaced by another cluster that fell between the "Samplers" and the "Strong Starters". The third MOOC, on the other hand, ran for eight weeks, but only included three assessments. In this case, the "Returners" and the "Mid-way Dropouts" were replaced with a cluster of "Samplers Who Comment", and by a much smaller cluster of those whose engagement was concentrated on the final week.

The authors then ask whether the engagement patterns identified in the four OU MOOCs are found in MOOCs having the same pattern by different universities, and whether engagement patterns are influenced by changes in learning design. The authors focus on five MOOCs from four institutions, a total of 32,942 learners: two long (7-8 weeks), one "talk only" with no assessment (6 weeks) and two short (3 weeks). It is worth mentioning that the authors had access to the data and time of the learner's first visit to a content step but did not have access to the date and time of any subsequent visits.

Their methodology to forming the clusters is as follows:

- Divide the data into weekly segments.
- For each course week, assign learners an activity score of 1 if they viewed content, 2 if they posted a comment, 4 if they submitted their assessment in a subsequent week, and 8 if they submitted it early or on time.
- apply the k-means clustering algorithm to split the learners into a small number of groups.

The study consisted of three phases based on the number of clusters for the k-means algorithm:

1. The dataset that was used included two long MOOCs, and the authors looked at clusters for which $k = 7$ provided the best fit.
2. The authors used the sets of data for "talk only" and two short MOOCs and explored why a value of 7 for k was not a good fit in these cases.
3. Finally, the authors used the most suitable value for k (3, 4, 5) for the three MOOCs datasets mentioned above, and analysed the results.

The resulting clusters of students in this research are presented in Table 2.

Table 2. The resulting clusters in (Ferguson et al., 2015).

Phase	Clusters
1	7 clusters from OU study (described in the introduction)
2	4 new clusters: ● **Surgers.** ● **Improvers.** ● **Saggers.** ● **Weak Starters.**
3	**TalkMOOC3** - 3 clusters: ● **Quiet.** ● **Week 1 Contributors.** ● **Consistent Engagers.** **ShortMOOC4** - 4 clusters: ● **Very Weak Starters.** ● **Strong Starters (Truncated).** ● **Returners (Truncated).** ● **Keen Completers (Truncated).** **ShortMOOC5** - 5 clusters: ● **Samplers (Truncated).** ● **Strong Starters (Truncated).** ● **Returners (Truncated).** ● **Improvers.** ● **Keen Completers (Truncated).**

To summarize, the authors conclude that "the results of a cluster analysis are dependent on the variables that are selected as significant by researchers" and that "a k-means analysis will produce k clusters for any value of k, but these will only be meaningful if priority is given to elements of the data that are significant in the context" (Ferguson et al., 2015, p. 81).

In the next paper "Portraying MOOCs Learners: a Clustering Experience Using Learning Analytics" (Khalil et al., 2016) the authors employ the same k-means clustering technique on a set of data collected during one of the courses on an Austrian MOOC platform - iMooX1. The course was active for ten weeks, with 838 participants across two groups - 459 internal university participants and 379 external volunteers. The clustering was done independently in both groups, university students and external participants, because the intentions of each group could vary.

The following variables were used to group the participants of the course:

- **Reading Frequency.** Indicates the number of times a user clicked on particular posts in the forum.
- **Writing Frequency.** Determines the number of written posts in the discussion forum.
- **Videos Watched.** Contains the total number of videos a user clicked.

- **Quiz Attempts.** Calculates the sum of attempts that have been spent on all ten quizzes.

The resulting clusters are the following:

Table 3. Case 1. The resulting clusters for university students in (Khalil et al., 2016).

Cluster	Description
"Dropout"	This group has low activity among the four variables. Only 10 students (out of 95) are certified, and the dropout rate is high.
"Perfect Students"	Most of the participants in this group completed the course successfully. This cluster is distinguishable by their videos' watching.
"Gamblers"	The certification rate was 94%. Both cluster 2 and cluster 3 share a high certification rate but differ in watching the videos.
"Sociable Students"	Smallest cluster, containing 4 students. Students in this cluster are the only ones that had been writing on the forums. The amount of certified students in cluster 4 totals to 50%.

Table 4. Case 2. The resulting clusters for external participants in (Khalil et al., 2016).

Cluster	Description
(No name was mentioned)	The certification rate of this group is 76.20%. The social activity and specifically reading in forums are moderate compared to the other clusters. Whilst the number of quiz trials is high.
"Perfect Students"	Holds only 8 participants. The certification rate in this group is 100%. Participants showed the highest number of written contributions and the highest reading frequency in the forum.

"Dropout"	This group showed a high dropout rate and a completion rate of only 1%.

All in all, the authors conclude, among other things, that "tomorrow's instructors have to think about the increase of the intrinsic motivation by those students who are only "playing the system" (Khalil et al., 2016, p. 276), suggesting that the instructors need to think about the ways to intervene when they spot low activity or elements of “playing the system” among their students. Additionally, the authors proclaim that "by analyzing the clusters, we think the opportunity to portray students' behaviours in the MOOC becomes possible nearby" (Khalil et al., 2016, p. 274), inclining that the clustering is a viable and realistic goal to pursue.

The next article “Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses” (Kizilcec et al., 2013) presents yet another research on grouping the students based on their activity. This time, “learners are classified based on their patterns of interaction with video lectures and assessments, the primary features of most MOOCs to date” (Kizilcec et al., 2013, p. 170). The authors use the data collected during three Computer Science MOOCs with around 97 thousand participants over nine assessment periods, nine weeks in total. The authors describe their main goal as “to strike a balance by identifying a small yet meaningful set of patterns of engagement and disengagement” (Kizilcec et al., 2013, p. 170). To achieve it, the authors define four learner trajectories as longitudinal patterns of engagement with the two primary features of the course – video lectures and assessments. The clusters are also compared with each other based on learner characteristics and behaviours.

The methodology consists of two parts:

1. Generate a rough description of each student's individual engagement in a course. For each assessment period, all participants are labelled:
    - “on track (T)” (did the assessment on time).
    - “behind (B)” (turned in the assessment late).
    - “auditing (A)” (did not do the assessment but engaged by watching a video or doing a quiz).
    - “out (O)” (did not participate in the course at all).
2. Apply the k-means clustering algorithm - the standard centroid-based clustering algorithm - to identify prototypical engagement patterns.

After following these steps, the authors present the resulting clusters of students, which are demonstrated in Table 5.

Table 5. The resulting clusters in (Kizilcec et al., 2013).

Cluster	Description
Competing	Learners who completed the majority of the assessments offered in the class. Though these participants varied in how well they performed on the assessment, they all at least attempted the assignments. This engagement pattern is most similar to a student in a traditional class.
Auditing	Learners who did assessments infrequently if at all and engaged instead by watching video lectures. Students in this cluster followed the course for the majority of its duration. No students in this cluster obtained course credit.
Disengaging	Learners who did assessments at the beginning of the course but then have a marked decrease in engagement (their engagement patterns look like Completing at the beginning of the course but then the student either disappears from the course entirely or sparsely watches video lectures). The moments at which the learners disengage differ, but it is generally in the first third of the class.
Sampling	Learners who watched video lectures for only one or two assessment periods (generally learners in this category watch just a single video). Though many learners “sample” at the beginning of the course, there are many others that briefly explore the material when the class is already fully underway.

To summarize, the authors mostly focus on giving course design recommendations and different suggestions about pedagogical aspects that should be kept in mind when creating a course. The authors also mention that they could identify work sessions if they would have used hourly time periods instead of weeks. Finally, the authors conclude that “learner patterns of engagement will change with time - a trend which could be explored through clustering engagement over present and future offerings of the same course” (Kizilcec et al., 2013, p. 176).

In the final paper “What Massive Open Online Course (MOOC) Stakeholders Can Learn From Learning Analytics?” (Khalil & Ebner, 2016) the authors carry out the research study about the development phases of a LA prototype and its integration into the MOOC platform called iMooX, which has been mentioned previously. The authors pose two research questions:

- “How can the Learning Analytics prototype trace students in a Massive Open Online Course Platforms?”

- “What are the patterns and revealed outcomes (evaluation) of applying Learning Analytics in MOOC platforms?”

The study is based on the data from two courses, which were active for 11 weeks total, with 1530 students participating in them. The data collected consisted of student activity traces regarding discussion forums, documents, videos and quizzes.

The methodology is based on the following three stages:
1. First step includes tracing the remnant touches of students, gathering their information, tidying and transforming the data, and storing their information securely in the server database.
2. Secondly, the student data is classified into categories of MOOCs indicators and after that, the data is analysed and visualized.
3. Finally, the results are inspected in order to reveal students' behaviour in courses as well as handing insights to MOOC stakeholders.

In a nutshell, the main way of forming clusters that the authors used is to count the number of events that each student participated in and then assign them to a cluster if they satisfy the criteria. The resulting clusters are presented in Table 6.

Table 6. The resulting clusters in (Khalil & Ebner, 2016).

Cluster	Description
Registrants	Students who enrol in one of the available courses
Active learners	Students who at least watch a video, post a thread in the discussion forums or attend a quiz
Completers	Those who successfully finish all the quizzes, but do not answer the evaluation form
Certified learners	Those who successfully finished all the course quizzes and reviewed their learning experience through the evaluation form

It can be concluded that the clustering in this paper is not based on the machine learning algorithm, but rather on participation in specific events. Therefore, this method may be beneficial for datasets with relatively small amounts of records. The method successfully produced several clusters that can be used for visualization and decision making, without utilizing any advanced LA method like Machine Learning.

2.7 Literature review conclusions

After conducting the review of the literature about the clustering of students based on their engagement level, it was discovered that in each particular case the resulting sets of clusters are quite different from each other, and it is not immediately obvious how to utilize this data to create efficient and meaningful visualizations. In other words, there is a problem of heterogeneity. Therefore, the sets of clusters that were identified previously should be inspected and analysed to understand how they can be made more abstract and dynamic, so that the process of forming the clusters becomes more flexible and universal. This will, for example, in theory allow instructors to decide how the clusters are formed based on their needs regardless of the course type or the number of students participating in the course.

Additionally, it was discovered that the k-means method of clustering is the most commonly used technique for forming activity groups of students based on their interactions with the course (Zhang et al., 2018; Ferguson et al., 2015; Khalil et al., 2016; Kizilcec et al., 2013). This method is well-established and allows for moderate customization, which makes it preferable for solving this type of task.

When looking at the clusters themselves, it was observed that the naming for them is arbitrary and is formulated by the authors themselves in each case. This might lead to different interpretations of similar clusters based on the name alone. For example, “Keen Completers” in (Ferguson et al., 2015) and “Perfect students” in (Khalil et al., 2016).

Finally, there is the problem of the dataset size. MOOCs can consist of a low number of students, transforming a Massive Open Online Course into a Small Private Online Course, which in-turn logically suggests the application of the advanced LA techniques like Machine Learning to be less preferable. In such a case creating a more abstract approach seems to be more favourable. This approach can consist of several standardized pre-defined clusters, which can also be adjusted by the instructors based on their requirements. Additionally, it was demonstrated in (Khalil & Ebner, 2016) that such an approach is possible and viable. Therefore, this study will be focused on this area and this theory will be used as a fundament for the next stages of the research.

Summary

This chapter has summarized the existing knowledge about the research topic and is crucial for the following research steps. By understanding the current state of the field, we can identify the areas, which can be expanded with new ideas and solutions.

Chapter 3

Open edX platform

This chapter provides a brief description of the Open edX platform, the structure of the data generated by it, and finally a description of the existing LA system, which is based on Open edX and which was used to evaluate the results of this research. The Open edX platform was specifically chosen due to the following reasons:

1. edX is widely adapted in Norway and it is currently used in 36 educational institutions.[2] UNIT, a directorate for ICT and joint services in higher education and research in Norway, also provides an Open edX platform, an open-source version of edX, to the interested parties, which makes it a useful tool for generating and analysing the educational data, especially in the context of Norway's educational environment.
2. Open edX is the main platform, on which the existing LA tool is built. Therefore, the artefact, development of which will be described in the following chapters, is also based on the same platform, since it will be integrated into the existing LA tool in order to evaluate its functionality.

3.1 Open edX

Open edX platform is an open-source version of the edX ecosystem. It was open sourced in 2013 after roughly a year has passed since the launch of the edX (Stanford News, 2013). The purpose of this platform, among other things, is to provide the educational institutions with the means to create their own analytics fit to solve concrete tasks. This means that the platform itself can be adjusted to the specific needs and that the tracking logs, which are generated during the course, and which contain information about users' activity, are also available for processing. This in-turn makes it possible to experiment with the collected data and extract meaningful information, which can then be utilized in many ways. For example, it can be used for creating different visualizations to overview the activity of the course and make decisions based on the presented information. The architecture of the Open edX platform is displayed in Figure 4.

[2] Directorate for ICT and joint services in higher education and research (UNIT). *Open edX-plattformen.* Retrieved from https://www.unit.no/en/node/572.

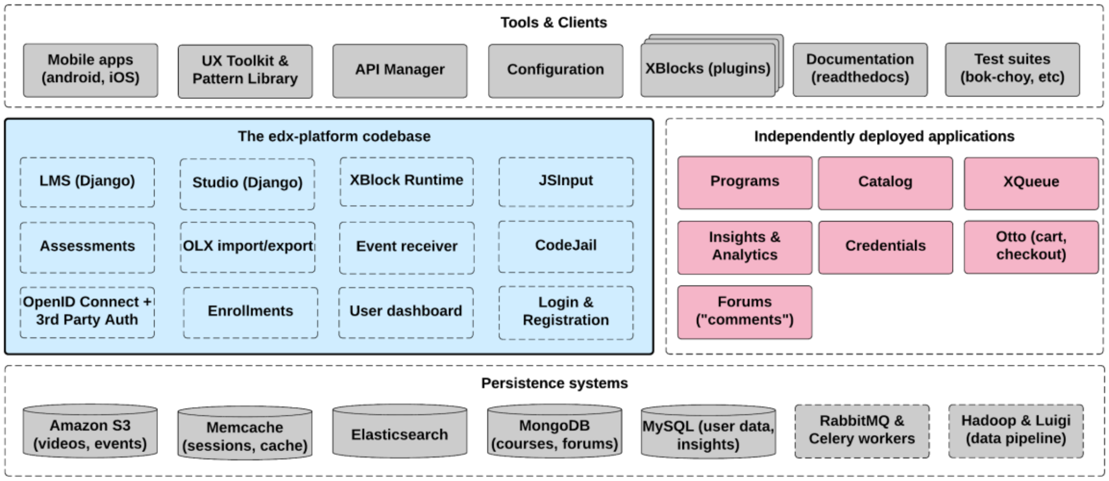

Figure 4. Open edX architecture.[3]

To better understand what data is being collected, its structure is briefly overviewed in the next part of this chapter.

3.2 Data and its structure

As it was described earlier, the data is generated by the Open edX platform based on the activity in the course. This data consists of several events that the system identified and saved in a JSON format. These events represent the interactions between the student and the system. For example, when the student navigates from one of the pages to another or starts watching the video. The example of just one event is presented in Figure 5. It shows one student's answer to a particular problem, as well as the evaluation of the correctness of the answer.

[3] edX. *Open edX Architecture.* Retrieved from https://edx.readthedocs.io/projects/edx-developer-guide/en/latest/architecture.html.

{"agent": "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/30.0.1599.101 Safari/537.36", "context": {"course_id": "edx/AN101/2014_T1", "module": {"display_name": "Multiple Choice Questions"}, "org_id": "edx", "user_id": 9999999}, "event": {"answers": {"i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_2_1": "yellow", "i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_4_1": ["choice_0", "choice_2"]}, "attempts": 1, "correct_map": {"i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_2_1": {"correctness": "incorrect", "hint": "", "hintmode": null, "msg": "", "npoints": null, "queuestate": null}, "i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_4_1": {"correctness": "correct", "hint": "", "hintmode": null, "msg": "", "npoints": null, "queuestate": null}}, "grade": 2, "max_grade": 3, "problem_id": "i4x://edx/AN101/problem/a0effb954cca4759994f1ac9e9434bf4", "state": {"correct_map": {}, "done": null, "input_state": {"i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_2_1": {}, "i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_4_1": {}}, "seed": 1, "student_answers": {}}, "submission": {"i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_2_1": {"answer": "yellow", "correct": false, "input_type": "optioninput", "question": "What color is the open ocean on a sunny day?", "response_type": "optionresponse", "variant": ""}, "i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_4_1": {"answer": ["a piano", "a guitar"], "correct": true, "input_type": "checkboxgroup", "question": "Which of the following are musical instruments?", "response_type": "choiceresponse", "variant": ""}}, "success": "incorrect"}, "event_source": "server", "event_type": "problem_check", "host": "precise64", "referer": "http:\/\/localhost:8001\/container\/i4x:\/\/edX\/DemoX\/vertical\/69dedd38233a46fc89e4d7b5e8da1bf4?action=new", "accept_language": "en-US,en;q=0.8","ip": "NN.N.N.N", "page": "x_module", "time": 2014-03-03T16:19:05.584523+00:00", "username": "AAAAAAAAAA"}

Figure 5. A sample of one Open edX event.[4]

In this form, it is hard to interpret this complexity, but after understanding the structure and consulting with the extensive event explanation available in the "EdX Research Guide"[4] it is possible to filter out most of the technical information and leave out only the important parts that are relevant to LA. For example, if we take the event in Figure 5, we can see that we do not really need the "ip" or "referer" fields, since they are purely technical and, most probably, do not provide useful information for LA. On the contrary, the "time", "event_type" and "event" fields are very relevant and can be used for describing, for example, the student's behaviour. These fields are important to note since they will be used for the main part of the study, namely the profiling algorithm.

3.3 Existing Learning Analytics tool - OXALIC

OXALIC is a LA tool designed to present an overview of the student activity using several different data processing and visualization techniques (Khalil & Belokrys, 2020). The main purpose of OXALIC is to provide different groups of stakeholders, mainly instructors and researchers, with useful representations of the data that is collected during the courses. This research focuses on

[4] edX. *Events in the Tracking Logs.* Retrieved from https://edx.readthedocs.io/projects/devdata/en/stable/internal_data_formats/tracking_logs.html.

one of the possible applications of this generated information, namely on grouping the students according to their activity and engagement in MOOCs. The resulting functionality is implemented as a module for the existing OXALIC platform.

The interface of the system is presented in Figure 6.

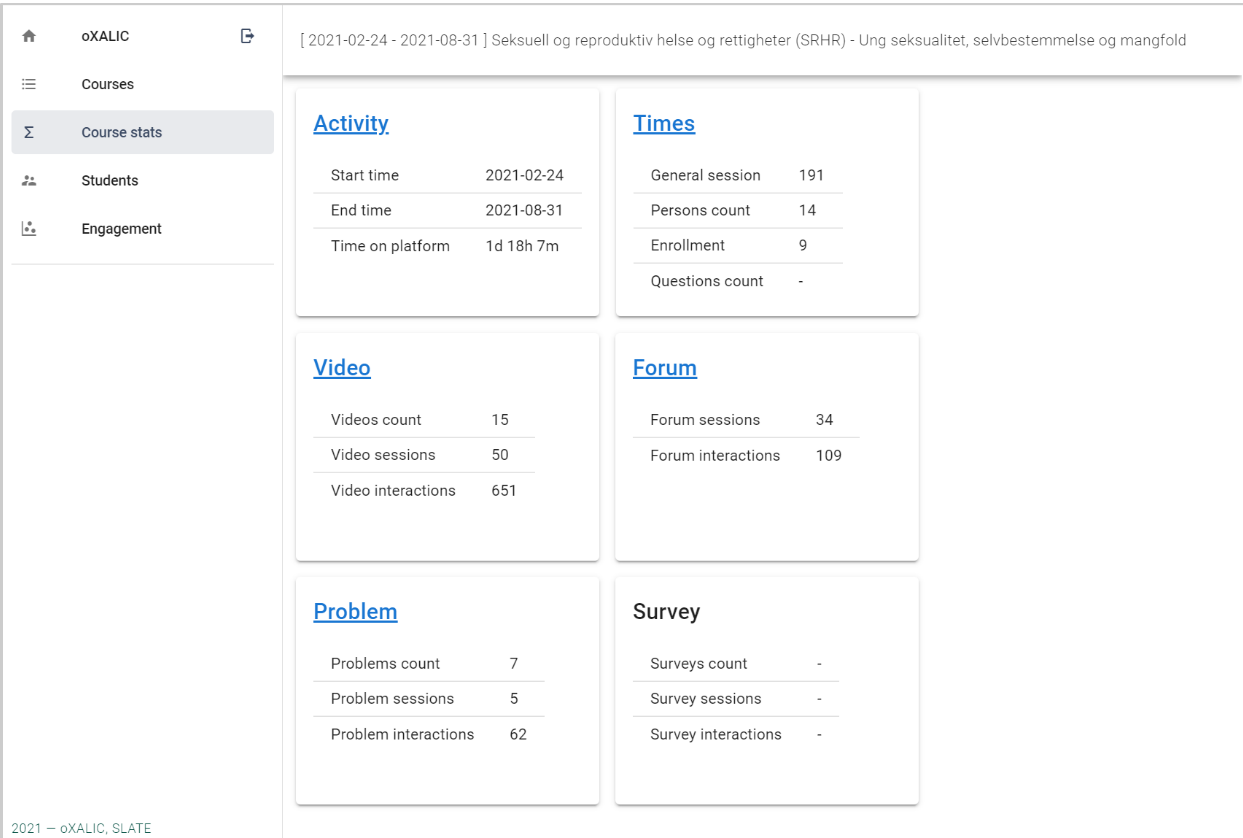

Figure 6. A course page in the OXALIC system.

The architecture of the tool is displayed in Figure 7.

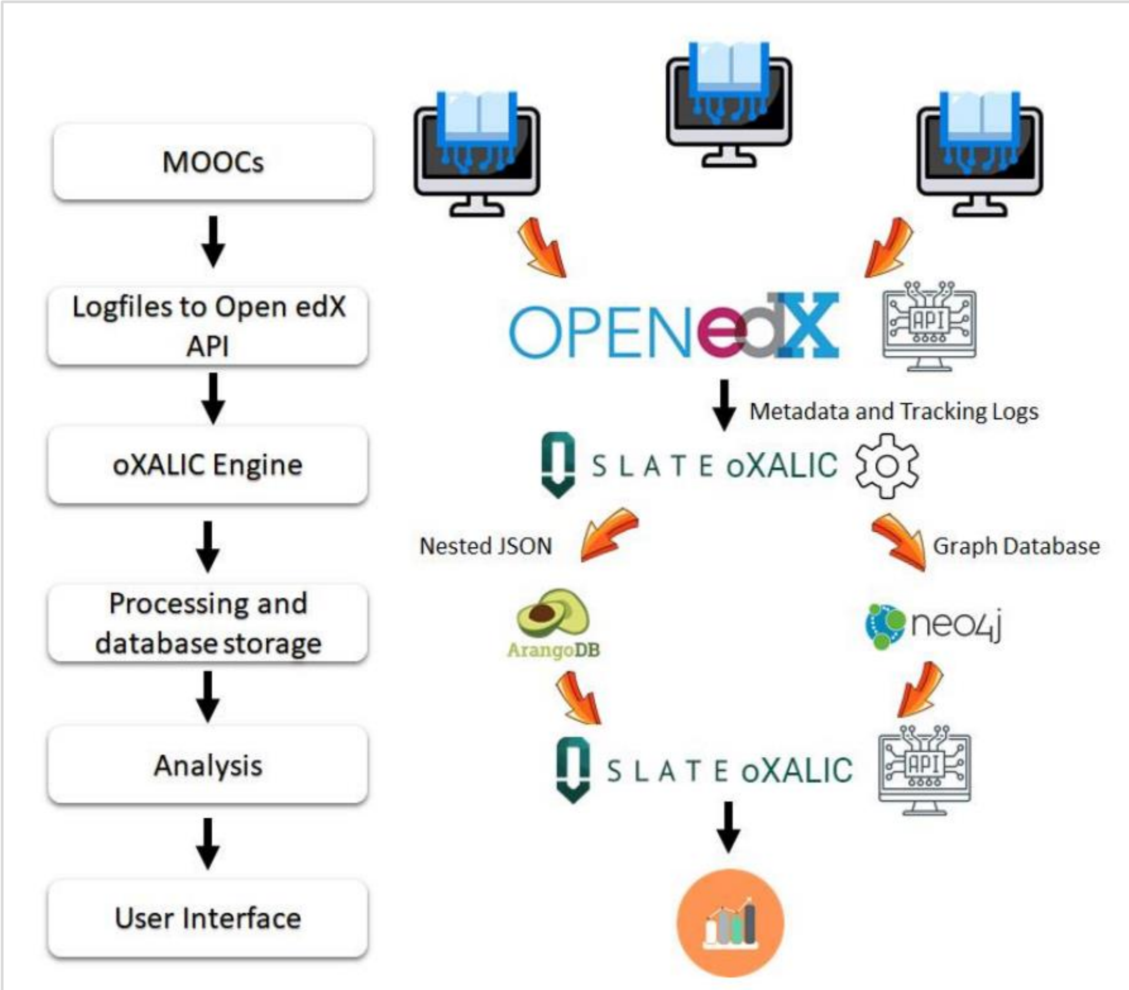

Figure 7. An overview of OXALIC architecture (Khalil & Belokrys, 2020).

The overview of the tool's architecture in Figure 7 represents the flow of the tracking data in the system. First, the tracking data is captured and transferred to the system. Then, this data is stored in the databases, one for generating the graphs, and the other for general processing of the data, which results in multiple pieces of information about the course. Finally, this information is formatted and presented to the users in a form of webpages with different statistics and graphs.

To understand the usefulness of the existing system, several groups of stakeholders can be identified:

- Instructors.
- Students.
- Course designers.
- Platform owners.
- Researchers.

The main group that benefits the most are the instructors. By using this tool, they can see students' progress and their interactions with different parts of the course. Based on this information instructors can guide students' progression and make interventions to help students improve their results. This is achieved by rich visualizations that are generated by processing the data collected during the current and previous course progression.

The second group are the students themselves. This can be achieved by providing a dashboard with aggregated personalized information about a student's course progression. Additional information can be presented as well, such as recommendations, predictions and different types of analysis that will help students to plan their education better and correct potential problems.

The third group are the course designers. By observing the information that is generated by the tool, the course designers can evaluate the efficiency of the course they have created so that the course can be improved for the next study period. This information can also be used for planning and creating new courses.

Platform owners can be considered as the potential benefactors as well. The data generated by the tool can be used to adjust the framework of the whole platform as well as the amount and types of the data that is being generated and stored. This way both the efficiency of analytics and data flows can be potentially improved.

Finally, the researchers can also use the data generated by this tool. For example, aggregated and filtered data. It can save time for researchers to receive information that was already filtered and refined based on the goals of the research, instead of executing the filtering and aggregating the data manually.

The analysis and the user interface parts of the architecture play the major role in providing the meaningful information to the stakeholders. This is achieved by presenting the tracking data in categories, which include the following:

- **Course stats.** This page provides a general overview of the course, with such information as the total number of students in the course, number of videos and other interactive parts, and several other pieces of information.
- **Forums.** This category contains the statistics about the interactions with forums, such as leaving a comment, searching, voting on someone’s comment, and other interactions.
- **Videos.** The video interaction analytics category presents multiple useful insights and has the bulk of the information generated by the tool. This is based on the fact that “videos are integral in MOOCs” (Khalil & Belokrys, 2020, p. 187), therefore they should be analysed the most. The generated information includes the amount of video plays, pauses and stops, the number of students, which played the video at least one time, total viewing time, and other observations. Individual videos can also be observed, with the amount of video interactions by students for each specific video. Finally, this information can be exported in multitude formats, such as PNG, PDF and CSV (Khalil & Belokrys, 2020).
- **Time spent on the platform.** This part provides an overview of the amount and length of the learning sessions in the course. A learning session is a period of time, when several

events are registered in the system one after another, with short pauses between them. This is a way of showing at what time of day and how long the student interacted with the course.

- **Events network.** Another piece of analytics is the network of the events. This network represents the chains of events, which happened in succession. This makes it possible to see which events represent the starting point for the student's interactions, and which events follow. For example, it can be concluded that the students almost always go to the video section after the first interaction, which might signify that the other parts of the course are less meaningful or noticeable for the students, and certain adjustments should be made in order to change that. The network is presented in Figure 8.

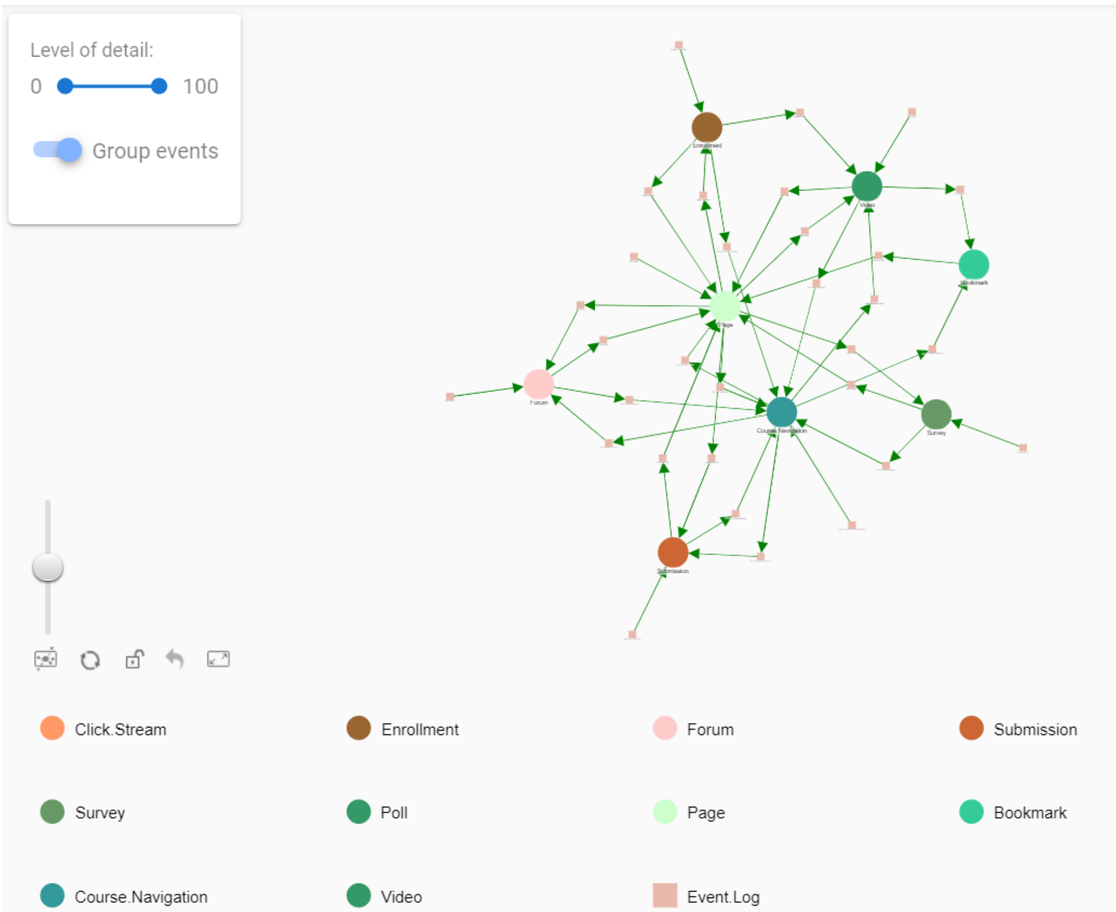

Figure 8. Events network in OXALIC (Khalil & Belokrys, 2020).

Summary

This chapter presented a brief overview of the Open edX platform and its architecture, alongside with the description of the existing LA tool, which is based on the Open edX tracking data. This allows us to understand how the Open edX platform works and what tracking data it can provide for the research purposes. Based on this information about the provided data, we can plan and develop the solution that transforms this tracking data into information for the researchers and instructors.

Chapter 4

Methodology and methods

In this chapter the main methods, which were used to work on the research project, are overviewed.

4.1 Overview of the research project stages

To help understand better the flow of the conducted research, as well as to have a clear picture overall of the whole project, an overview of the steps taken during the work on this research is presented in Figure 9.

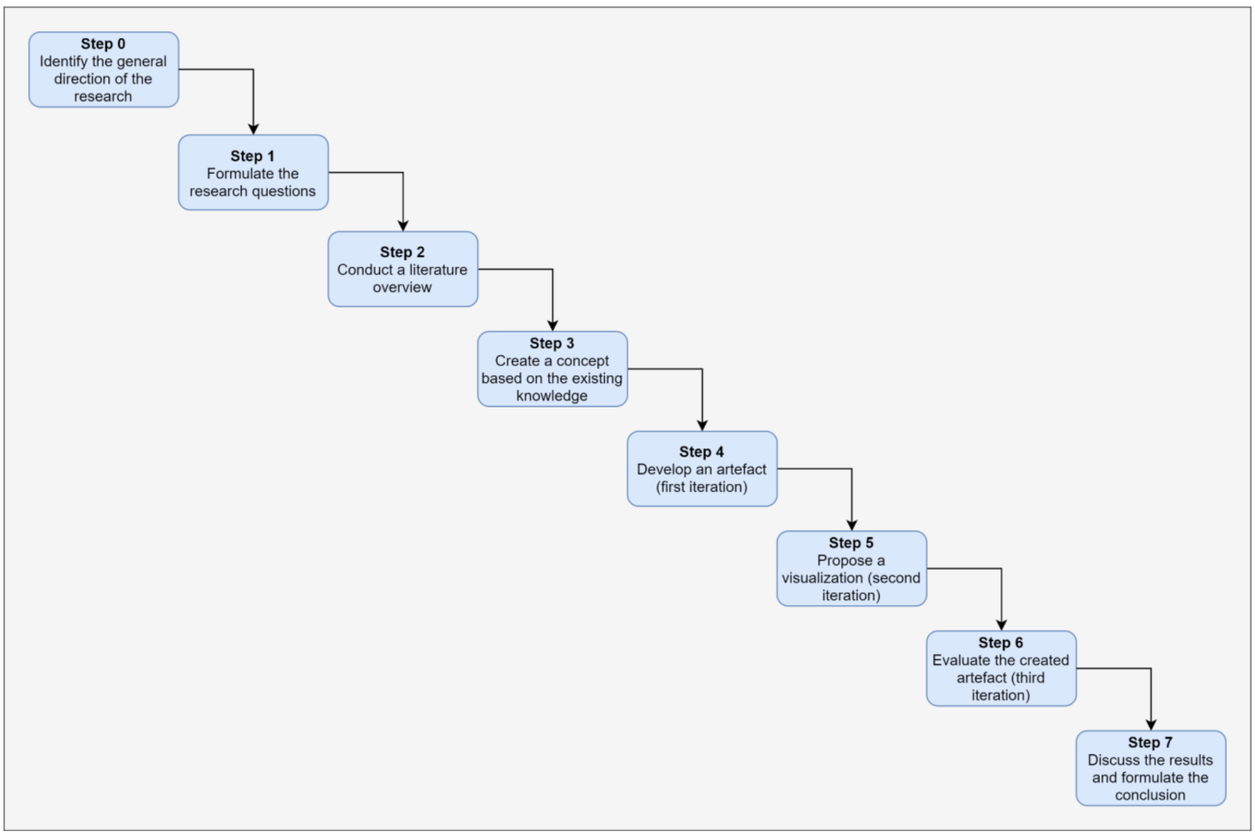

Figure 9. An overview of the steps taken during the research project.

The whole project is partitioned into several steps. Each step can be shortly described as follows:

Step 0. The general direction for the research was chosen. On this step the general area of research was chosen from many available topics and the relationship with the supervisors was established.

Step 1. The research questions were formulated. During this step, a more concrete direction for the research was outlined by understanding the specific parts of the research area, which can be expanded with new knowledge.

Step 2. The literature review was conducted. By using the research questions from the previous step, a few articles were identified and analysed, and the resulting information was noted for the future steps. Additionally, this step allowed me to understand the current state of the research area and plan the next steps.

Step 3. The concept for the chosen research area was created. The concept was based on the existing knowledge and used to answer the research questions, thus contributing to the existing knowledge.

Step 4. The artefact was developed based on the created concept. This step includes the practical part of the project, i.e., the programming and testing of the artefact, as well as the creation of the infrastructure to support its main functionality.

Step 5. The visualization part for the created artefact was proposed. Since the artefact at this point did not have any way of presenting the results to the user, a way of visualizing said results was described and demonstrated.

Step 6. The created artefact was evaluated based on the user feedback and interactions. After the users had some experience working with the artefact, their feedback was collected and analysed. Resulted information will be used in the following step.

Step 7. The results of the research and the evaluation were discussed. At this point, there is enough information to answer the posed research questions based on the development of the artefact and users' feedback. The conclusions were also presented alongside with the remarks about the potential future work on the research subject.

After describing the whole process step-by-step, each step then is described in more detail in the following chapters.

4.2 Design science research

The main system of methods, which was chosen for this study, is design science research. “Design science research is a method that establishes and operationalizes research when the desired goal is an artifact or a recommendation” (Dresch et al., 2015, p. 67). The goal of this kind of research is to discover new knowledge by solving practical problems. It means that by analysing a particular area in the field, a task of solving a specific issue or achieving a specific goal can be formulated. And by successfully creating a working artefact that solves the identified issue or goal, the researcher contributes to the knowledge base, thus expanding it and allowing for further additions. Additionally, this methodology is oriented on both theoretical and practical applications of the existing knowledge, which allows the results of the research to contribute to two fields of study - the creation of the theoretical foundations and practical solutions that are based on these foundations (Dresch et al., 2015).

4.3 Three cycle view

The general description of the design science research methodology is presented in Figure 10.

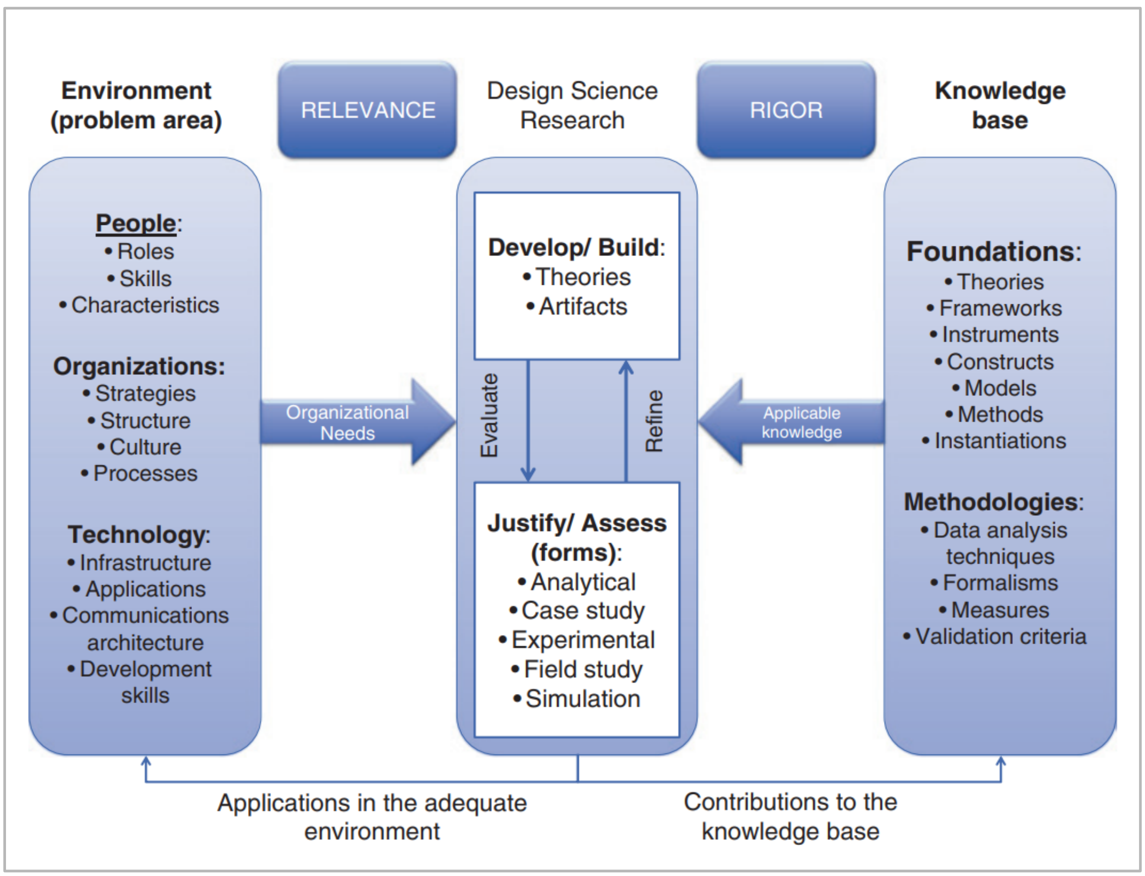

Figure 10. Overview of the “Design science research” methodology (Dresch et al., 2015).

The methodology is based on two sources - the environment and the knowledge base. These two sources provide two main parts, on which the methodology is based - relevance and rigor. Let us investigate them in more detail.

The “relevance” part consists of the principle that there exists a certain environment that contains several relevant problems, which need to be solved. These problems arise inside the organizational environments, which include different technologies and people with different levels of skill, thus formulating the specific goals for the study based on the design science research methodology. For this part, solutions resulting from conducting such study are considered the main contribution. Such solutions can be used and further developed by the organizations, justifying the relevance of the whole study (Dresch et al., 2015).

The “knowledge base” part includes all the existing knowledge and known methodologies regarding the chosen research field. For such a practical oriented approach as design science research, the existing knowledge usually consists of already existing solutions and artefacts, which are used as a basis for further research (Dresch et al., 2015). Since the goal of any research is contribution of new knowledge, the existing knowledge base is often not sufficient for the development of the new artefacts. Therefore, for the studies based on the design science research methodology, the researchers often choose to try new ideas and experiment during the development of the solution for the problem (Dresch et al., 2015). Such an approach allows new knowledge to naturally be created in the process of study, which is then contributed to the existing knowledge base.

By identifying the organizational needs and analysing the current state of the knowledge, the next part of the study begins. This part consists of two phases - development and assessment. The artefact is firstly developed based on the organizational needs and existing knowledge, then evaluated using the methods, which are established and accepted by the scientific community. A few improvements and corrections are then formulated, which are then used to improve the existing artefact. This iterative process can be executed several times until it satisfies the organizational requirements, culminating into the working solution, development cycle of which adds new insights to the knowledge base.

4.4 Design science research principles

To help the scientists, who choose the design science research as the methodology for their studies, to direct, assess and adjust their work, Hevner et al. (2004) in their article “Design science in information systems research” formulates seven important methodological principles. These principles provide useful guidelines, which can be considered crucial for conducting a successful research project (Dresch et al., 2015). The principles are presented in Figure 11.

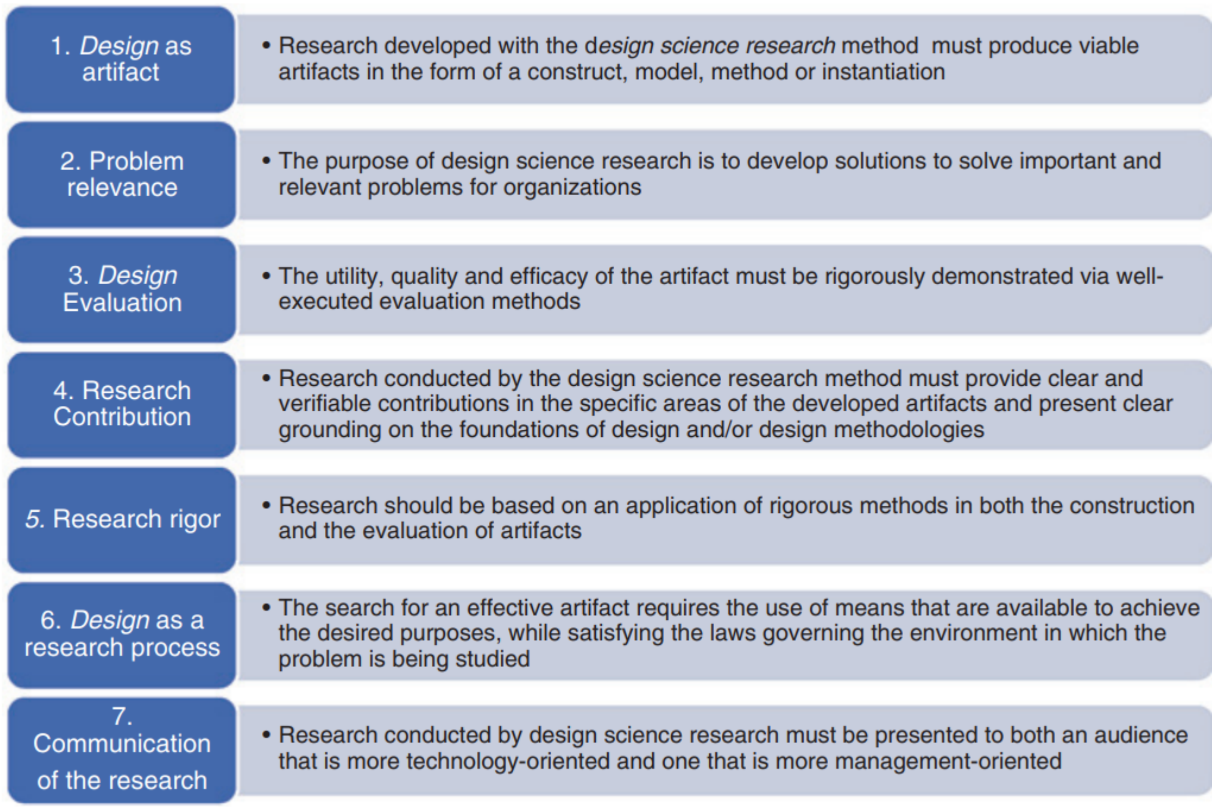

Figure 11. Seven principles of the "Design science research" methodology (Dresch et al., 2015).

The first principle "Design as artifact" conveys that for the research to be successful, some sort of an artefact should be created. That is, the theoretical knowledge should be applied to solve a practical problem, which results in a working solution.

The second principle "Problem relevance" insists that the solution should be focused on resolving existing and relevant problems, which require immediate attention. Naturally, by creating an artefact that addresses a certain task, other researchers can use it for their own studies, which may even expand on the newly created artefact.

The third principle "Design evaluation" states that the artefact should be evaluated in order to justify its relevance and usefulness. Without this step, it is not possible to reliably understand if the goals, which the artefact is aimed to reach, were actually reached. Without this, the knowledge generated during the research cannot be considered good enough for it to be contributed to the knowledge base.

The fourth principle "Research contribution" implies that the created artefact should clearly show, which areas of the knowledge base are being contributed to, and what these contributions are. That is, the research should answer specific research questions, which are, alongside with the description of the development process, the main additions to existing knowledge.

The fifth principle “Research rigor” suggests that the development and the evaluation of the artefact should use the well-established and accepted scientific methods and techniques. This is especially important for the evaluation part, since the usage of a number of widely accepted evaluation methodologies guarantees heterogeneous results, which can be reliably compared to each other.

The sixth principle “Design as a research process” conveys that the study should be conducted as a process, during which different techniques and methods are used, to satisfy the organizational needs. These techniques should be within the established problem area to keep the focus of the study on solving the specific task.

The seventh principle “Communication of the research” is about presenting the results of the study to the main stakeholders. This allows the accumulated knowledge to be shared, and additional adjustments and corrections to be formulated, which ultimately leads to creating a better solution, which satisfies all the interested parties of the research.

By following these principles, the research project should result in a successful achievement of set goals. The exact results of the application of the described methodology and the aforementioned principles are presented in Chapter 8.

4.5 Evaluation and its purpose

After deciding which methods should be applied for conducting the current study, it is also important to choose the methodology, which will be used for evaluating the results of the study.

The evaluation is one of the crucial stages in the development of the artefact based on the design science research methodology. This is supported by the authors of the article “Design science in the information systems discipline: an introduction to the special issue on design science research” (March & Story, 2008), which is then presented by Dresch et al. (2015, p. 70) as that “the developed artifacts should be properly evaluated in terms of their utility and viability to demonstrate their practical and academic validity”. Indeed, this statement can also be deduced by simple logic: if the artefact is properly evaluated by the experts, it can be determined if it brings any value to the field, thus the development of this artifact contributes new knowledge. Additionally, even if the artefact does not provide substantial value to the interested parties, the process of its development, alongside the evaluation stage, already provide knowledge for other researchers, who are interested in the same field. It means that even if the evaluation results show that the artefact is not fulfilling its purpose in some way, this fact in and of itself brings knowledge into the research area.

Two approaches to conducting the evaluation can be considered - quantitative and qualitative. Quantitative approach focuses on the numbers and statistical analysis of the existing data, while the qualitative approach targets the meaning of the events, their origins and how they influence the present. Considering the current research and the state of the artefact, a qualitative approach is more beneficial in getting meaningful evaluation results. This choice is mainly justified by the fact that no existing statistical data regarding the use of this specific artefact is present. It means that in order to evaluate the newly created features, they need to be explored and experienced, i.e., evaluated, by the users, which in-turn results in useful feedback. This feedback is considered the result of the evaluation, and it is used for further development and improvement. Additionally, this feedback is also valuable for the other researchers, since they will not need to spend the time on conducting such evaluations, they will get the results straight away.

There are many different evaluation methods available for the purpose of this study: surveys, questionnaires, interviews, and several other ones. The next part describes, which evaluation method has been chosen and how this method will be used and applied to the developed artefact.

4.6 Evaluation method

For the purpose of establishing the evaluation method for the artefact the book “Social research methods” by Bryman (2016) has been chosen, namely the part about structured interviews and self-administered surveys, as well as the general rules that should be followed when formulating the questions for any evaluation.

The author states that the rules to how the questions in social science studies should be formed, are well known and have been so in many years. However, “it is one of the commonest areas for making mistakes” (Bryman, 2016, p. 264). To address this, the author proposes three simple rules of thumb, which should be used for the purpose of social science research. These rules are as follows:

- **Always bear in mind your research.** The questions should always be relevant to the research and its goals, in one way or another. That is, each question should provide at least some piece of information, which can be used in the study.
- **What do you want to know?** When formulating questions, it is important to remember that they should not be too general or too abstract. For evaluating the artefact, it is always beneficial to focus on specific parts, such as functionality, user interface, visualizations, and so on. This way each part of the artefact can be thoroughly evaluated and adjusted accordingly, based on the provided feedback.
- **How would you answer it?** It is also useful to try and answer your own questions by pretending to be the person you want to get the feedback from. By doing so, it is possible to correct or rephrase the questions early, so that they focus more clearly on the problem area.

Keeping these rules of thumb in mind, the main way of evaluation now needs to be chosen. Two possibilities were considered, namely the self-administered survey and structured interview. The choice fell on the latter, and it was based on the following arguments:

- Although the self-administered surveys can be considered more convenient for the respondents, as well as overall quicker and easier to manage, the time constraints do not allow for waiting for the answers from the respondents.
- Additionally, this kind of evaluation does not allow for quick clarifications if the answer does not fully cover the area, which is being evaluated. In order to get more information, several rounds of surveys will be required.
- Finally, the interviewees agreed to the structured interview and considered it a convenient way for evaluating the artefact.

After the method for the evaluation has been outlined, the questions for the structured interview should be created. In order to do that, several specific recommendations for formulating the questions were used, which are also mentioned in (Bryman, 2016). These recommendations suggest avoiding ambiguity, technical terms, leading questions, as well as double-barreled questions, such as “What do you think about A and B?”, which essentially split the attention of the respondent into two problem areas, instead of one. Finally, it is also recommended to avoid questions, which include negatives, such as word “not”, since this can confuse the respondent and their understanding of the inquiry will be the opposite of what it was aimed at originally.

By following the aforementioned rules of thumb, as well as the more specific recommendations, a set of questions was formulated, which is presented in Appendix A. Results of the evaluation are described in Chapter 7.

Summary

In this chapter the main methodology and methods have been described in detail. It is important to choose and understand the methodology for the research before conducting the main part of the study for it to be successful and fruitful. This step can be considered similar to having a plan for the main bulk of the work, i.e., how the things should be done in order to get meaningful results.

Chapter 5

Learning Analytics algorithms

This chapter focuses on the overview of the profiling algorithm, as well as the visualization for this algorithm.

5.1 Profiling concept

In this part the profiling concept is described, which was created based on the conducted literature review, and the initial requirements to the artefact.

5.1.1 Motivation

Based on the review of the literature regarding the task of grouping students according to the activity data collected from the online course, it can be concluded that the main clustering technique that is used is k-means clustering, alongside several other techniques. The issue in our case is that we do not have a big enough dataset to apply these Machine Learning techniques effectively and get meaningful results in the process. Therefore, one of the possible solutions in this situation is to create a custom grouping technique, which uses the basic activity data to produce clusters of students grouping them by the level of engagement. The distinct feature of this algorithm is that the results can be dynamically adjusted based on the preferences of the instructor. This means that the instructor is able to choose the weights for the specific activities, which might reflect the engagement better in the given course.

5.1.2 Main profiling concept

As it was mentioned earlier, the tracking data, which is generated by the Open edX platform, contains the interactions between the student and the online course. These interactions come in the form of events. Each event belongs to a certain category, for example different server-side events, video interaction events, forum interaction events, etc.[4] For the sake of the proposed concept, only a select number of event types are chosen, since these specific events represent the main bulk of the actual activity between the student and the system. It means that the events like video interactions and forum interactions are preferred over the server-side events and different network-related events, which are purely technical, and do not represent the students' activity.

The event groups that are used for grouping are as follows:

- Login event
- Video interaction events
- Problem interaction events
- Forum interaction events
- Surveys and polls events

For each day, each student gets a score based on the events and activities they have participated in. If the student participated in one of the aforementioned activities at least once - they get a score for this type of activity, which is based on the weight assigned to this activity. The final score is a sum of scores for each event type that the student participated in during the day. This score can then be summed up again to form a weekly or a monthly activity score.

After generating the score for each student, the presence of outliers is checked. If there are students with too low or too high scores - they are placed in their own respective groups - very low participation and very high participation.

After removing the outliers, the average score is calculated among the students that are left. This score will indicate the average engagement level of students.

Finally, the students are grouped as low engaged, moderately engaged and highly engaged by having low, moderate and high scores respectfully.

To have a more precise control over the groups of students, the instructor can change the weights of the events manually or specify that some of the events are more important than others to avoid direct manipulation of numbers that represent weights. Additionally, the algorithm can be further expanded by the addition of the following feature - the instructor can change the average score that is used as a threshold for grouping by using the statistics from previous periods, similar courses or their own perception of how engaged the student should be in their course to be considered active or not. Currently, this functionality is not yet implemented due to the lack of the historical course data containing average scores of the students from the past courses.

5.1.3 Profiling algorithm

The proposed profiling algorithm is depicted in Figure 12.

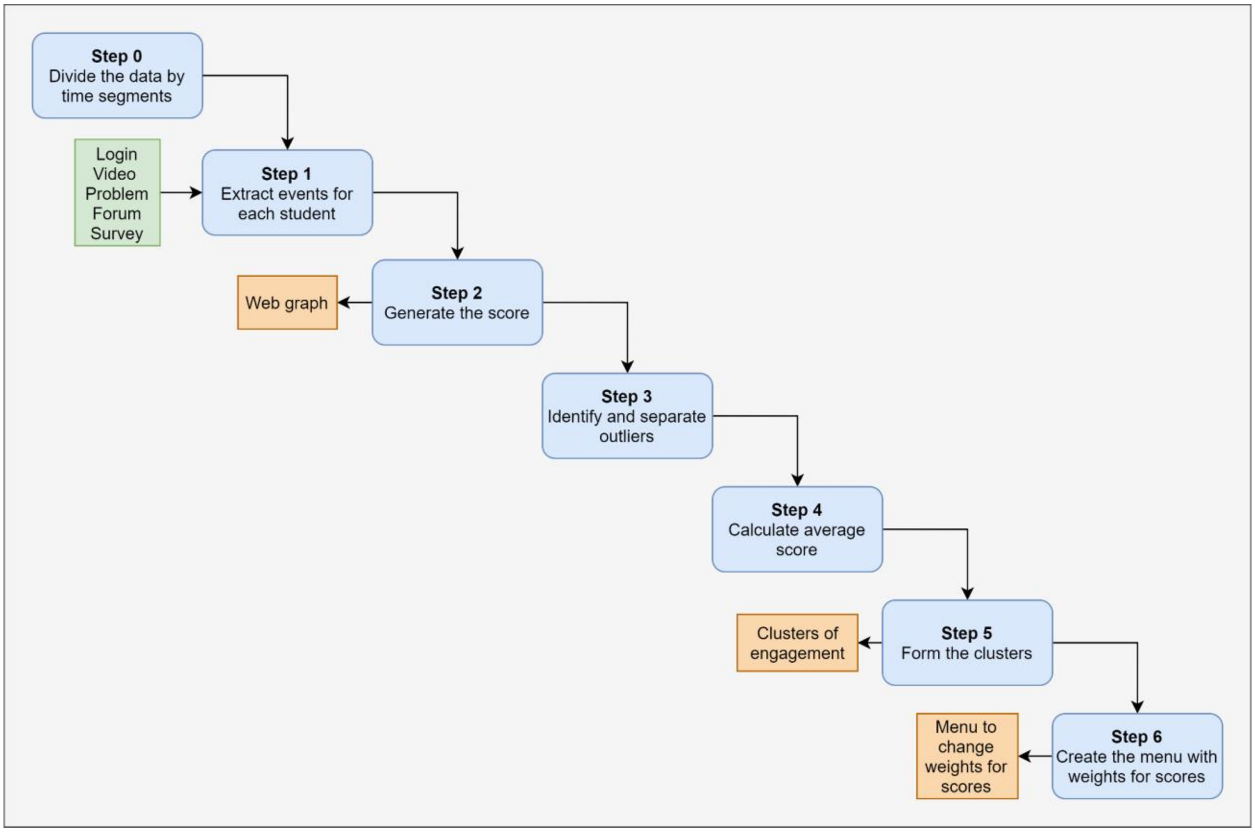

Figure 12. An overview of the profiling algorithm.

Step 0

First of all, the time period is established in which the events are happening. For this purpose, a simple metric is used, which registers if the events happened during one day at least once. For example, a student played a video today.

Step 1

After that, the events are extracted from the tracking data for each student in a form of variables. For each day, the number of events according to the select four groups is counted and saved.

Step 2

Each variable has a weight assigned to it to allow for more precise grouping. The default value of the weight is 1.0. For each day, for each student in the course, a score is generated, which consists of a sum of the event scores that happened during this day. For example, if the student has viewed a video and visited a forum during the day, and assuming the weights have the default value, then the final activity score for the student for this particular day will be 2.0, that is 1.0 for video interaction, 0.0 for problem interaction, 1.0 for forum interaction and 0.0 for survey interaction. Therefore, each student now has a score for each day that represents their activity. Next, these scores can be summed up to form the weekly and monthly score for each student.

Step 3
The outliers are identified at this point. Students with the score of 0.0 or the ideal score, i.e., the highest score among all the students, are placed into separate groups.

Step 4
Next, the average score is calculated based on the total scores of each student for a selected period of time. This period can be selected by the instructor depending on their needs, but as of the moment of writing, only the whole course period is used.

Step 5
Then, the students are grouped into engagement clusters based on the proximity to the average scores. These scores are represented by two thresholds - low activity threshold and high activity threshold. Students, whose scores are lower than the low activity threshold are placed into a low engagement group, while the students with scores higher than the high activity threshold are placed into a high engagement group. Finally, the students with the scores between the low and high activity thresholds are placed into the moderate engagement group.

Step 6
A menu that controls the weights of each event is available to the instructor as well. By using this menu, the instructor can choose one type of activity and see the engagement of the students based on the selected activity type. For example, when the instructor wants to see how engaged the students are based on the interactions with the videos in the course. Additionally, the instructor can determine custom value for weights, to make certain types of events more influential when forming the engagement scores. For example, the instructor can assign a higher weight to forum interactions over the videos, because forum activity is more important in this particular course, and it is more representative of the students' engagement.

Result
As a result, the algorithm produces five clusters, or groups, of students - no engagement, low engagement, moderate engagement, high engagement and highest engagement. In general, the main bulk of students will be positioned in the low, moderate and high engagement groups, while the outliers will go to no and highest engagement groups.

The implementation of the presented algorithm should be sufficient for answering the proposed research questions, namely, how to identify the engagement based on the Open edX tracking data, as well as what is required for achieving this. Additionally, the resulting information generated by the algorithm should allow for extra features like web graphs and typical clusters. Web graphs represent activity of each student in a form of a web for a period of time divided by the events the student participated in. Typical clusters can be formed according to the clusters identified in the literature review.

5.2 Visualization concept

In this part the visualization concept for the proposed profiling algorithm, which has been described in the previous part, is presented.

5.2.1 Main visualization concept

After the algorithm successfully processes the tracking data, the activity for different periods of time is calculated in a form of variables. These variables are stored in the arrays and can provide knowledge to the skilled experts, who can analyse this information by directly accessing the variables storage. However, this way of conveying information is not suitable for the general public. Additionally, it is not feasible for the instructor to go through arrays of variables to try and extract any useful knowledge that they can then apply in their daily work. Therefore, some sort of visualization is crucial in order to make the solution actually useful.

A concept of such visualization is presented in Figure 13.

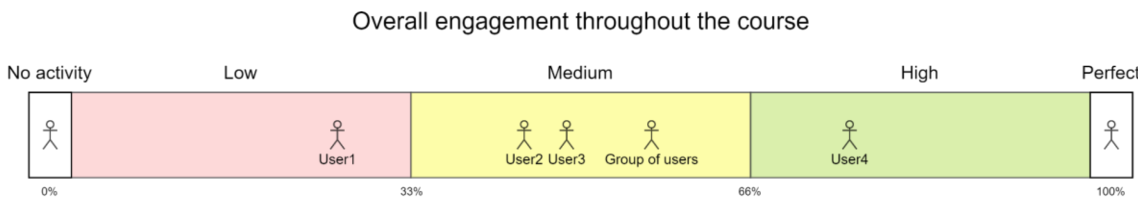

Figure 13. An overview of the visualization for the profiling algorithm.

This visualization consists of five arbitrary groups of students - “No activity”, “Low”, “Medium”, “High” and “Perfect”. Each group represents the activity of the students during a certain period of time. At the moment of writing, this period spans over the whole course length, but this period can be customized by adding this feature to the existing functionality. After calculating the scores for the users, they are placed in the respective groups and displayed for the instructor, providing a clear overview over the activity of the students in the given course. The groups of users are created if there are several users present with similar activity. These groups can be interacted with and inspected in more detail to get an overview of each student in such a group. This grouping is required to keep the visualization clear and prevent cluttering. The group content is presented in Figure 14.

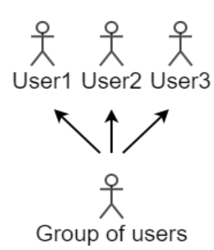

Figure 14. A group of students with similar engagement.

As it was mentioned earlier, the engagement can also be based on a certain period of time. For example, the visualization that represents the engagement for one week for the whole group of students, or just for one student, is presented in Figure 15.

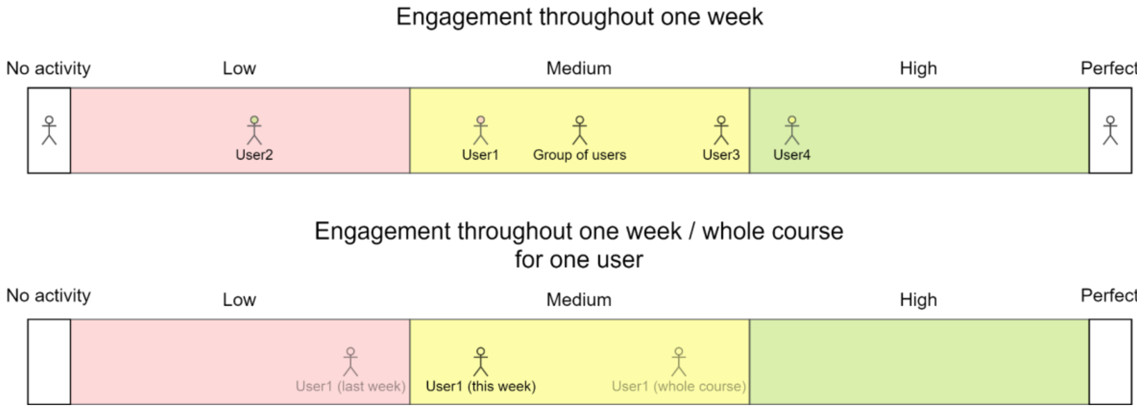

Figure 15. An overview of the engagement throughout one study week.

The notable thing that can be mentioned for this particular visualization is that the student icons have a colour assigned to them that represents their activity throughout the course. This way the issue of a misleading engagement representation can be prevented. That is, a student might have very high engagement during the first week of the course and complete all the main activities several weeks in advance and have low engagement during the second and third week. It means that without any additional method of distinguishing students based on their past activities, this particular student will be placed into the low engagement group during the second and third week, which can be misleading about their overall activity in the course, since they were very active in the first week and completed all the assignments for the following weeks. The colour assigned to the student that represents their overall activity solves this problem and allows the instructor to adequately assess the student's engagement in the course.

Finally, to have a more detailed and practical representation of the engagement clusters, the instructor can choose an alternative representation, which consists of a list of all students who are assigned to the corresponding groups. This representation is displayed in Figure 16.

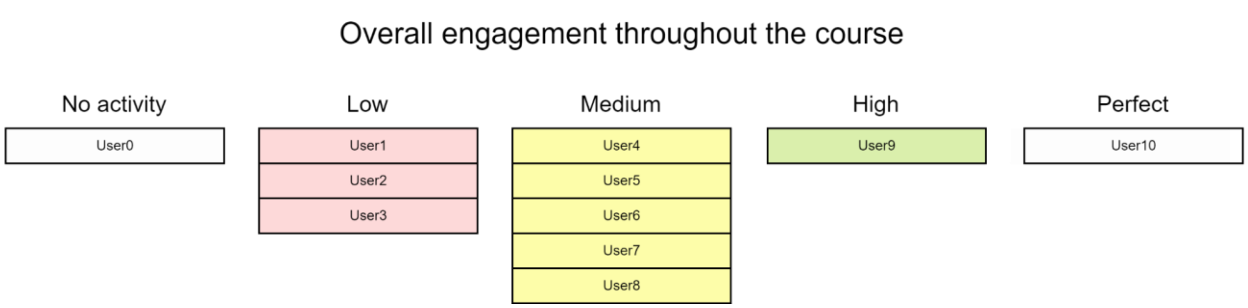

Figure 16. An alternative representation of the students' engagement.

This representation can be considered more practical since it displays all the students in the course in one table, without making groups of students with similar activity.

Summary

This chapter provided a detailed description of the proposed algorithm. This is a theoretical basis for the actual practical implementation, which is presented in one of the following chapters. It also displayed a concept for how the generated information can be presented to the instructor in a meaningful way. It is not enough to have a working algorithm without any way of presenting the results of its work, therefore some sort of visualization is crucial.

Chapter 6

Artefact development

This chapter presents the process of implementing the profiling algorithm in code, as well as creating the visualization for this algorithm.

6.1 Formulating the requirements

In order to outline the specific direction for the development of the artefact, a set of requirements should be first established. These requirements define what specific goals the resulting artefact should reach. They consist of two categories - functional and non-functional requirements. Functional requirements focus on specific aspects of the artefact, such as certain functionality or methods of processing the data. Non-functional requirements describe the perceived properties of the artefact, such as the responsiveness of the interactive elements and positive user experience in general, when interacting with the interface of the tool.

Functional requirements:

- The tracking data should be stored in the ArangoDB database.
- The tracking data should be processed and formatted in a specific way, so that it can be used by the grouping algorithm.
- Groups of activity should be generated and stored in variables and act as a foundation for the visualizations.
- Multiple types of visualizations should be used to display the groups of students.
- The events in the tracking data should be grouped into day-long periods of time, to allow the activity to be presented in future based on different time periods.

Non-functional requirements:

- The interactive elements of the tool should be robust and intuitive to use.
- The visualization should be easily readable by the users of the tool.
- The visualizations should convey the generated information in the code in such a way, so that it becomes meaningful by the common user.
- The visualizations should use neutral colors and look pleasant in general.
- The user should be able to always understand, which part of the tool they are currently using, and how to switch to other parts.

The requirements will be used as a guideline during the implementation of the artefact. By satisfying these requirements, the core functionality of the artefact can be successfully created, which will allow for further additions and corrections.

6.2 Algorithm development

In this section the development process of the proposed profiling algorithm is described.

6.2.1 General information

According to the “Design science research” methodology, the development of the artefact consists of two main stages - building of the solution itself and, subsequently, assessing the resulting artefact. This whole process is iterative as well, meaning that this cycle of development and evaluation can repeat multiple times, until the desired result is reached.

The development of the artefact for this research therefore includes two parts - creation of the artefact itself, which is described in this and the following part, and then its evaluation, the results of which are presented in Chapter 7. Due to the nature of the task, being the implementation of not only the profiling algorithm itself, but the visualization for the generated results as well, the first part of the artefact development therefore combines two smaller parts - implementation of the algorithm, described in this part, and the creation of the visualization, presented in the following part. This way the whole process can be separated into smaller bits for easier understanding and exposure of more details regarding each part of the development.

6.2.2 Open edX platform and tracking data

Before the main part of the artefact development begins, we need to better understand the structure of the tracking data, collected during the interactions between the student and the system, as well as the way this data is stored and how it can be processed for the purpose of this research.

As mentioned in Chapter 3, when a student interacts with a course based on the Open edX platform, a piece of tracking data is captured and stored on the storage medium. This data is recorded in JavaScript Object Notation (JSON) format, which is a lightweight way of storing data, and which is easy-to-process by both humans and machines alike (Json.org, 2002). This tracking data contains multiple fields, which in-turn contain different values representing the observations about the student’s interaction with the system. This includes, among other things, the username of the student in the system, the time the interaction has happened, the type of the event, and many other pieces of data. Not all these pieces can be used for this research since they do not immediately present any value for the study. For example, such fields as “agent”, “host”, “session”, and several other ones, are aimed at capturing certain technical information, which can, of course, potentially be used for the sake of this research, but it was decided to skip it and only focus on the most meaningful tracking data.

After understanding how the tracking data is generated and stored, we can move to the next step and begin the implementation of the proposed algorithm for the grouping of students.

6.2.3 Tools and technologies used for development

The artefact consists of two major parts - so-called back-end and front-end. The back-end part covers the processing of the data and different calculations and manipulations with the data. The back-end part further consists of two smaller parts - the database and the functions, which process the data from the database. The front-end part is aimed at presenting the results of the processing of the data by the algorithms in the back-end part to the user and will be covered in the following sub-chapter.

The database provider, which was chosen for the development of the artefact, is ArangoDB. This is an open-source NoSQL database, which has flexible models for traditional "key: value" formatted data, as well as graphs and documents.[5] It also supports its own declarative query language, ArangoDB Query Language (AQL), which can be considered an advantage over other database providers, such as MongoDB, since it allows for more complex queries with multiple access patterns in one query.[6]

For processing the data after it has been stored in the database, a programming language should therefore be selected. There are many programming languages to choose from for the development of such an artefact, but since the artefact consists of both back-end and front-end parts, the choice fell on one of the most popular programming languages at the time of writing - JavaScript. It can handle both processing of the data, as well as presenting the results using one of the modern front-end frameworks (Mozilla, 2021). The front-end part will be described in more detail in the following sub-chapters.

Next, an integrated development environment (IDE) should be chosen in order to start writing the code that implements the described algorithm. For this purpose, Visual Studio Code was used. This IDE is suitable for many different programming languages, but it is considered to be one of the best specifically for developing solutions based on JavaScript programming language. The IDE supports several modern technologies, like syntax-highlighting with autocomplete feature, built-in GIT functionality and the ability to install optional extensions and packages to expand the existing features of the IDE.[7]

---

[5] ArangoDB. *Key Features, ArangoDB Documentation.* Retrieved from https://www.arangodb.com/docs/stable/.
[6] ArangoDB. *What you can't do with MongoDB.* Retrieved from https://www.arangodb.com/solutions/comparisons/arangodb-vs-mongodb/.
[7] Microsoft. *Visual Studio Code Documentation*. Retrieved from https://code.visualstudio.com/docs.

After establishing the core tools and technologies for the development, the actual implementation of the algorithm can be started. This process is described in the following sub-chapter.

6.2.4 Development process

The implementation of the artefact can be presented in a step-by-step overview of each part of the algorithm. This whole process can be described as follows:

Step 0. After the data is uploaded to the database, an automated script is executed, which adds additional fields with the information about the year, week and day of the week. These variables represent the date when the event happened and can be used for separating the events into time periods. At the time of writing, this functionality is not utilized since the grouping happens for the whole course, without the ability to group it based on dynamic time periods, which are selected by the instructor.

Step 1. The tracking data is stored in the database as a collection of events. Each event represents an interaction between the system and the user. Events can also be server-side, that is, they are emitted by the server and do not represent an actual interaction. An example of an event is presented in Figure 17.


```
▼ object {18}
    username : sysdrift
    event_source : server
    name : edx.course.enrollment.activated
    accept_language : no
    time : 2018-09-10T18:57:04.408175+00:00
    agent : Mozilla/5.0 (Macintosh; Intel Mac OS X 10.13; rv:61.0) Gecko/20100101 Firefox/61.0
    page : null
    host : srhr.mooc.no
    session : 0b8a709e0f361e7f17b61af1c901baf5
    referer : https://srhr.mooc.no:18010/home/
  ▶ context {4}
    ip   : 10.5.0.93
  ▶ event {3}
    event_type : edx.course.enrollment.activated
    id   : 4effc8754387b8eb502228cad891798b
    year : 2018
    dayofweek : 1
    week : 37
```


Figure 17. An example of a record of a single tracking event in the database.

To start working with the data, it needs to be queried from the database for further processing. To achieve that, the connection with the database is first established, and then the information is collected by making an AQL query. The query is displayed in Figure 18. This query filters the fields, which were deemed to be irrelevant for this research, and then groups the events by user.

```
FOR log IN tracking_logs
  FILTER log.context.course_id == ${this.courseName}
  SORT log.context.course_id, log.context.user_id, log.time
  COLLECT user_id = log.context.user_id, username = log.username INTO groups = {
    "event_type": log.event_type,
    "time": log.time,
    "week": log.week,
    "day_of_week": log.dayofweek
  }

  RETURN {
    "user_id": user_id,
    "username": username,
    "user_log": groups
  }
```

Figure 18. A query that collects the required information from the database.

The resulting array of information contains all the users in the course, together with all their respective events, which happened throughout the course.

Step 2. After the tracking data has been collected and processed, the activity score should be calculated. It is achieved in the following two stages:

1. For each student, the amount of the selected tracking events is counted. These selected event groups, as it was mentioned earlier, are video events, problem events, forum events and survey events. As a result, each user now has information about how many events from the aforementioned event groups happened on each day of the course.
2. Next, each student receives an activity score for the whole course, and for each week in the course. The score is calculated by summing the associated weights for each type of the event for each day of the course. For example, if the student watched a video and left a comment on the forum but have not interacted in any way with problems and surveys, and the weights have their default values of 1.0, the student will have a score of 2.0 for this particular day - 0.0 for no activity with problems and surveys, 1.0 for interacting with the video and 1.0 for participating in forum activities.

After this step, each user has an activity score assigned for them for two periods of time - for the whole course and for each week in the course.

Step 3. Next, the outliers are placed into activity groups based on the score, which was calculated in the previous step. During this stage, the outliers are identified and separated into the respective groups - no activity and highest activity. Student(s), whose score equals zero, are placed in the first group, and student(s) with the highest score are placed into the highest activity group. As a result, we have placed the outliers into the respective clusters.

Step 4. The thresholds for the grouping are calculated at this point. First, the lowest and the highest activity scores are identified among all the students for the period of time. Then, the low and high activity thresholds are calculated. Low activity threshold equals to highest activity score minus lowest activity score, divided by three. High activity threshold equals low activity threshold multiplied by two. After this step is done, we have two thresholds for separating the students into three groups of activity.

Step 5. The final activity groups are formed during this step. By using the thresholds, which are calculated in the previous step, each student is placed into one of three clusters - low activity, moderate activity and high activity. If the student's activity score is lower than the low activity threshold, they are placed into the low activity group. Subsequently, if the student's activity score is between the low and high activity thresholds, they are placed into the moderate activity group, and if their activity is higher than the high activity threshold, they are naturally placed into the high activity group.

After executing all the described steps, five clusters are formed as a result. Each student is now placed into one of the activity clusters - no activity, low activity, moderate activity, high activity and highest activity. This can be considered as a successful implementation of the proposed algorithm, and the development of the first iteration of the artefact is finished.

6.2.5 Challenges

Naturally, multiple challenges were encountered during the development process. The most significant challenges, as well as the attempts to overcome them or at least mitigate their influence on the results of the research, are described below:

- **Lack of metadata.** Each online course that uses Open edX as the platform has several metadata fields assigned to it. These fields contain such information as the total number of videos in the course, the name of the course, the number of participants on the course, and several other pieces of information. Some of this metadata is generated automatically by the platform software itself, while the other needs to be filled in by the course instructor or maintainer. During the development of the artefact, it was discovered that some of this information was missing in the course's metadata. Therefore, it was either skipped, decreasing the ability to get a more whole picture of the course, or it was manually calculated using custom scripts to get the required information. This approach, while it solves the task, is not ideal since the script is not as reliable as the course instructor manually filling in the metadata about the course. To prevent this from happening, it is suggested to require the course instructors or persons responsible for managing the course to fill in the metadata page before, or shortly after the course starts.
- **Disparity in platforms for video content.** Another issue, which was discovered during the analysis of the tracking data, is that some of the courses used different platforms to

host video content. Since the tracking logs do not contain any technical information about the videos, such as video duration, tags, and other pieces of metadata, the information required for more in-depth analysis of the video interaction events is received by making Application Programming Interface (API) calls. This type of analysis is not currently used in the artefact, but the existing LA tool OXALIC already uses this information about video interaction to create additional insights for the course instructor. However, since the API calls are not universal, they need to be adjusted for each particular course in order to get the metadata for the videos in the course. In certain cases, there is no API publicly available for a particular platform, and in this case, there is no way of forming and presenting any meaningful observations about the interactions between the students and video content in the course. Such observation is, for example, which part of the video is most often interacted with among the students, which can mean that this part contains the most useful information for the students, and that it can be condensed into a video summary for quicker access. To solve the issue of disparate platforms for video content a good suggestion will be to use one platform for all the videos in the course, and make sure that the platform has a publicly available API.

- **Different tracking event IDs.** To identify the type of events in tracking logs, each event record has a field named “event_type”. This field contains one of the key words, which explicitly identifies the type of the event. However, due to different versions of Open edX platform itself, as well as the type of the platform, on which the course is interacted with - web-browser or Android app, the information in this field can differ for the same type of event. It means that, for example, that the event in tracking logs in the Android app will have a certain value in the field “event_type”, while the same event in a web browser tracking log will have different value in this field. This challenge was overcome by including all types of the events on all platforms and most recent versions of the Open edX platform.
- **Inconsistency in activity score thresholds.** In order to separate the students into different groups based on their activity, which is the final result of the algorithm, the thresholds for the activity are calculated. Initially, these thresholds were calculated by using average values of activity scores between all students. However, with this approach, the groups leaned too much into the extremes, i.e., most of the students were either in the high activity group or the low activity group. This issue was fixed by applying the logarithmic distribution to group students. Each activity score is converted into a different logarithmic value, which is then used to put the student into one of the activity groups. The activity thresholds were also changed to logarithmic values to support the updated distribution method. This way the students are now more evenly distributed between the activity groups, which in-turn allows for better visual representation of the student activity.
- **Lack of actual tracking data.** In order to apply traditional ML techniques and extract meaningful insights, the amount of data available for the analysis should be sufficient. Since the tracking data generated by the Open edX platform contains personal information, such as the student's name and IP address, which can be used to identify the physical location of the student, the process of acquiring the tracking data can take several months. Additionally, the number of students in each course varies greatly, from as few as 10 students to as many as 200 and more. These conditions make it difficult to create a consistent and reliable way of processing the tracking data and presenting consistently

useful and meaningful results, which are independent from the aforementioned shortcomings. To overcome this challenge, a custom grouping algorithm was created instead of a method based on ML. This way, the lack of tracking data and the patterns in it should not prevent the results to be generated and presented to the instructor.

6.3 Visualization development

In this part the process of creating the visualization for the proposed algorithm is overviewed.

6.3.1 General description

After finishing the back-end part for the artefact, that is the calculation of the users' activity, this information now needs to be presented to the user in some convenient and meaningful way. In order to achieve that, the second part of the artefact was developed, which consists of the webpage with the dashboard. This dashboard uses the calculated scores and weights for activities to present the engagement groups to the instructor. They can then interact with this dashboard by selecting the type of activity they want to investigate more closely, and change the weights to make certain types of activities prioritized over the others, altering the distribution of the students between the activity groups based on the more relevant activities.

The development of the visualization part consists of two smaller parts. The first part is the original rough implementation, which can be considered as a testing ground for different types of visualizations and experiments with positioning of the interface elements. The second part is as close to the visualization concept as possible, with the inclusion of one alternative visualization in the form of a Sankey graph.

6.3.2 Tools and technologies used for development

In order to create the visualization for the artefact, a framework with functionality, which supports such type of presentation of importation, is required. However, since the artefact is considered to be a part of the existing LA tool OXALIC, it needs to use the framework of this tool. Therefore, before starting the development, this framework was inspected and researched. Prior to this point, no experience with such a framework was present and it was acquired through practice and online courses from the ground-up.

The framework, which is used in OXALIC, is called "Vue". This is one of the three most well-known front-end frameworks, other ones being "Angular" and "React" (The State of JavaScript, 2021). The framework has extensive documentation and examples of working implementations[8],

[8] Vue.js. *Introduction.* Retrieved from https://vuejs.org/v2/guide/.

which were used during the development of the front-end part of the artefact. One of the main features of “Vue” is the component-based architecture. It means that the different parts of the one application can be separated into multiple components, which can be combined in one place, and still retain the ability to communicate with each other. This is very reminiscent of the object-oriented programming paradigm, which focuses, among other things, on dividing the complex functionality into smaller bits, making them easier to understand and manage.

Additionally, to extend the existing set of user interface components, which are available in Vue, an additional library of visual elements is also used in OXALIC. This library is called “Vuetify” (Vuetify, 2021) and it is based on the “Material Design”[9] design language, which is being developed by Google. One of the core components of this specification is the “card” concept, which allows separating the information into highly customizable and interactive visual blocks. This makes it easier to organize the content and present it to the users in a convenient and intuitive manner.

The next step, which follows the outlining of the tools and technologies for the implementation of the front-end part of the artefact, is the description of the development process itself, and the challenges, which were encountered along the way.

6.3.3 Development process

The development of the front-end part for the artefact included two stages. The first stage was aimed at creating a rough implementation, which contains all the features from the concept, but with the basic visual representation. This way it is possible to test the functionality and adjust parts of the tool on-the-fly, without worrying about the overall page composition and how the elements are positioned on the page. The resulting view is displayed in Figure 19.

[9] Google. *Material Design Homepage*. Retrieved from https://material.io/.

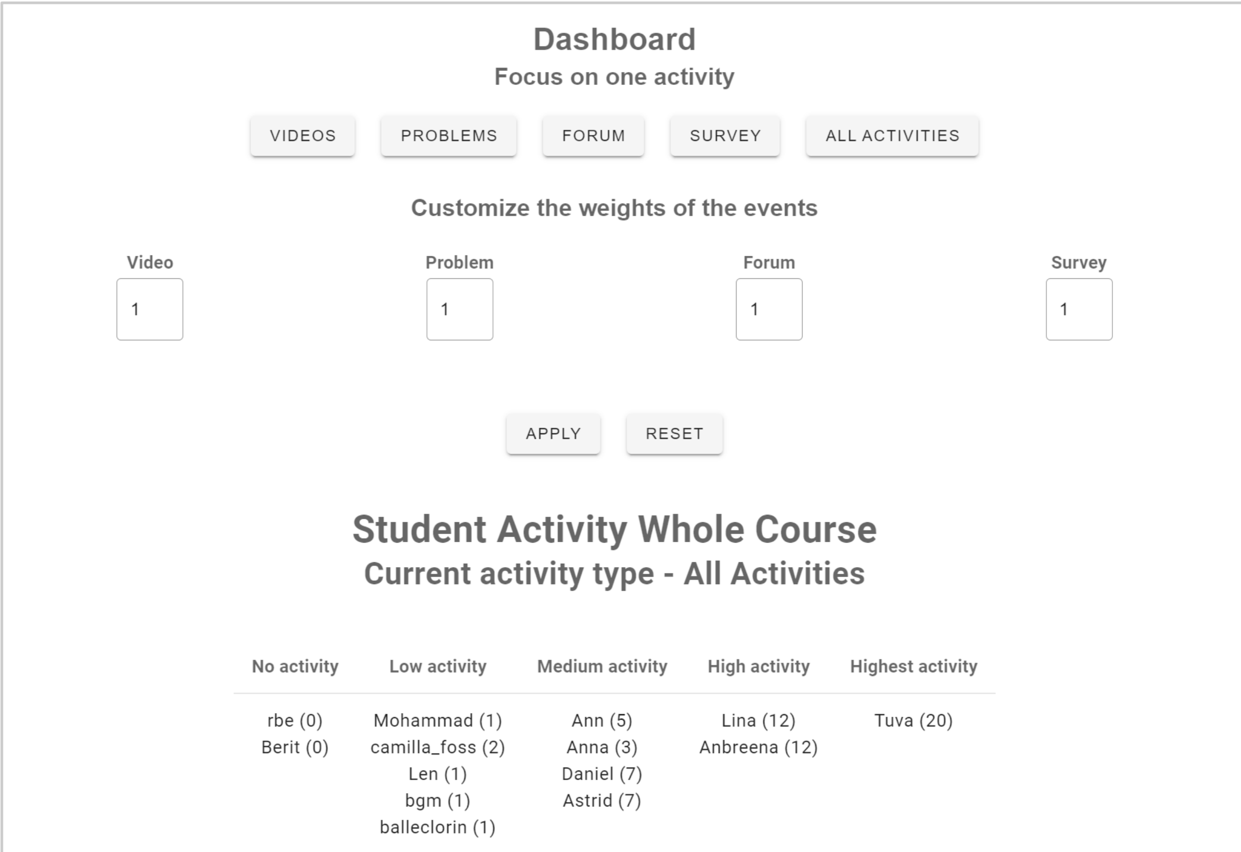

Figure 19. The first implementation of the visualization.

The page consists of three parts, which can be described as follows:

- The upper part contains several buttons, which allow changing the weights for the different types of activity, making it possible to focus on just one activity with the click of a button. Functionally, each button changes the weights of the other three activities to zero, while setting the weight of the chosen activity to one, that is 100%.
- The middle part has four fields for manual input of the weights for each type of activity. This can be useful, for example, to make several types of activities less important than others, by changing their weights to a number, which is lower than one, i.e., making it less than 100%.
- The lower part represents the main goal of the visualization - the groups of activity. Since each user has an activity score associated with them, they are therefore placed into their own group. Users with no activity are clustered in the “No activity” group, and so on. Groups can also be empty if there are no students with high enough activity scores. The numbers in the parenthesis are the activity scores for each user, which were present in the visualization for debugging purposes.

After testing the functionality and making sure that everything works and there is enough data to form the visualization, the second, more refined interface was implemented, see Figure 20.

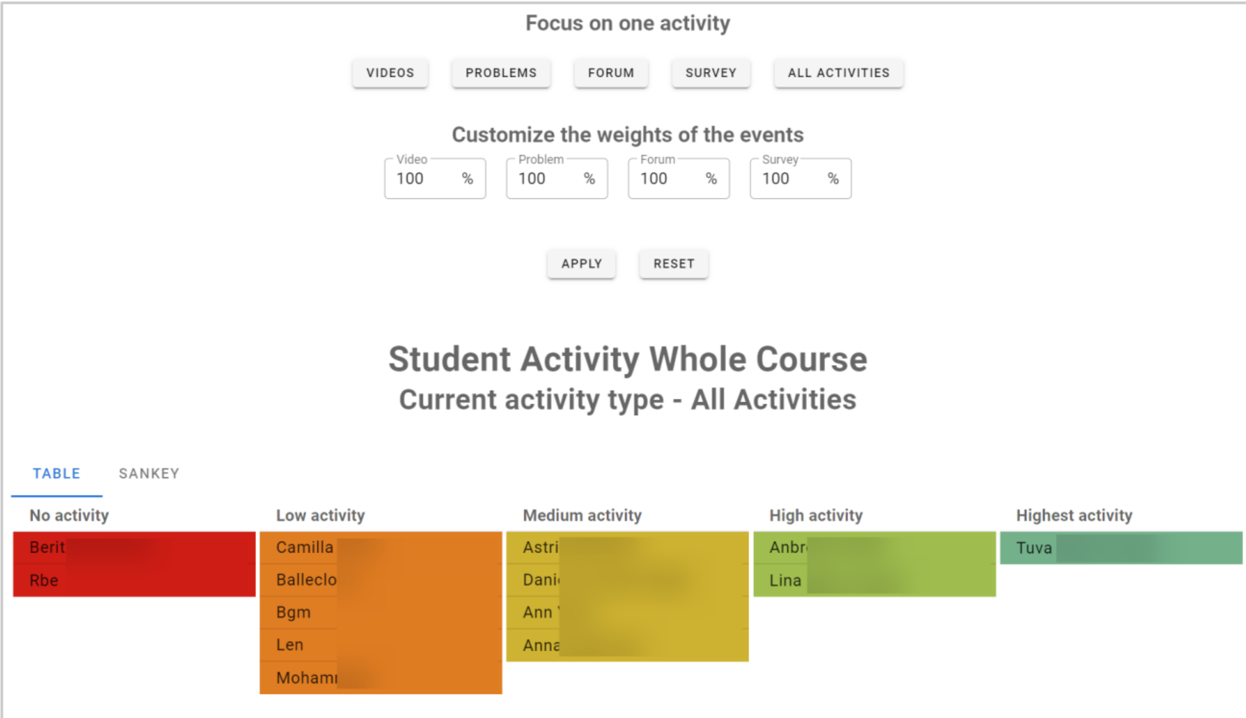

Figure 20. The final implementation of the visualization.

The main changes in comparison with the previous implementation are the following:

- The values of input fields were converted to be percentage-based and capped at 100% to prevent unnecessary overcomplication. This change makes the values more intuitive and easier to understand in contrast with the float-based numbers.
- The activity scores in the table view were removed since they were used only for debugging and no longer required to be displayed at this stage. It was also found that they might cause additional confusion, since their purpose and the way they are calculated are not obvious to the user.
- Finally, each group now has colour representation based on the activity of the student. This feature allows the users to quickly grasp the meaning of the grouping, since the colouring is intuitive and starts with red for “No activity” and ends with green for “Highest activity”.

Additionally, a second visualization was developed in order to present an alternative to the default view. This is also useful for evaluation, since the two representations can be compared and it can be decided, which is more preferable from a point of view of the actual user. In order to add an additional visual element, a visualization grammar G2[10] was used, which expands the available visualizations with more advanced and more scalable charts. After analysing the graphs, which are available in the grammar, a Sankey chart was chosen as an addition to the default table view. The resulting chart is presented in Figure 21.

[10] AntV. *G2, a Visualization Grammar*. Retrieved from https://g2.antv.vision/en.

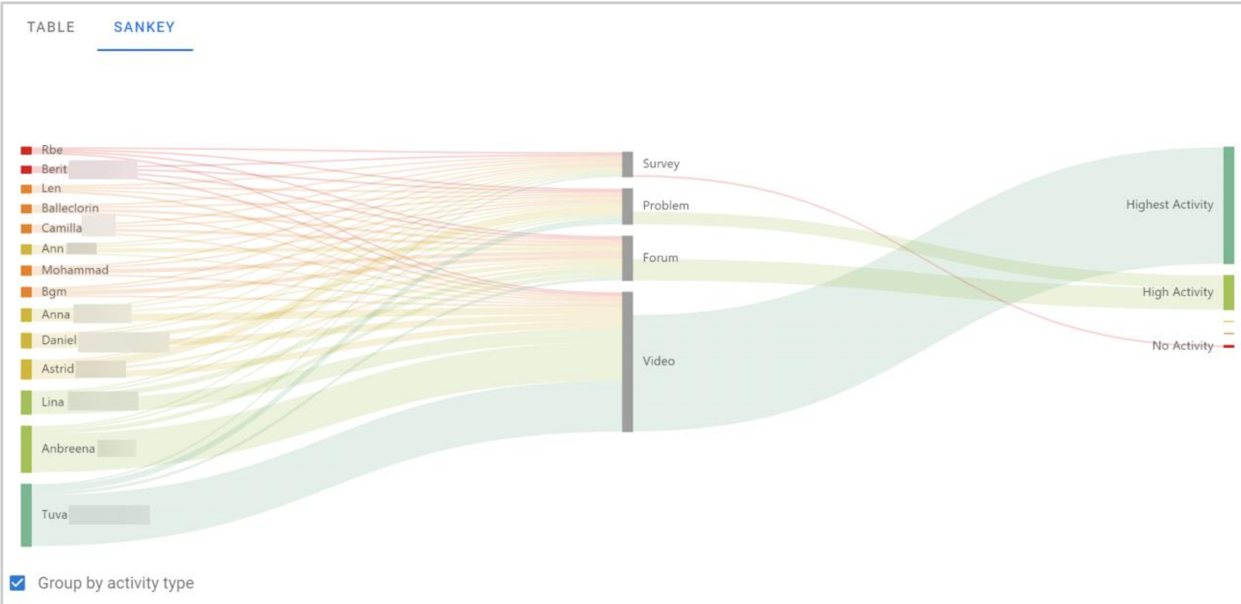

Figure 21. The Sankey chart implementation.

The Sankey chart visualization contains a more detailed view on the types of activity, in which the students are engaged during the course. The distribution of student activities can also be summarized, that is which activities in the course are most engaged with, and which are under-represented. Overall, it can be said that this is an alternative visualization, which contains more information than the table view, but in a different form. Therefore, these two visualizations can be considered eligible for a fair comparison in the evaluation stage.

Due to time constraints, the bar representation, which was described in the concept, is considered as a part of the future work. It will take one full development cycle to prototype, implement, test and evaluate said visualization, which is not possible with the current time limit.

During the implementation of this part of the artefact, no major challenges were encountered, apart from constant prototyping and polishing the visual composition. However, one challenge can still be mentioned, that being the problem of scalability. It was discovered that the implemented visualization works well for the courses with a relatively small number of students. When trying to apply the same visualization to a larger course, the perceivable usefulness was partly diminished by the fact that it takes more time to analyse the table and alternative Sankey chart. Naturally, the more students are present in the course, the longer it takes to scroll through the table, or the Sankey chart. The potential fix for this particular issue could be to implement a third representation, which will specifically address the problem of visualizing many students at once. This issue is covered in Chapter 9, where other potential additions through future work are also described.

As a result, the artefact is now in a fully working state, with both back-end and front-end parts successfully implemented. The next step is the evaluation stage, during which the feedback from the actual users, that is the instructors, will be collected and analysed in order to formulate a list of adjustments and fixes for the next iteration of the development process.

Summary

This chapter presented the overview of the practical implementation process of the proposed theoretical concept using the modern technologies and tools, alongside with the challenges encountered in the process. The implementation of the algorithm is vital for understanding its usefulness when working with real-world data. This part is also indispensable for the evaluation stage. Additionally, this chapter summarizes the details about the visualization development process. It is an essential foundation for answering the research questions of the thesis.

Chapter 7

Evaluation

This chapter describes how the profiling and visualization parts of the artefact were evaluated, and the results of said evaluation are presented.

7.1 Evaluation process

The evaluation consisted of a structured interview with two course instructors, who used the main LA tool, OXALIC for over 9 months. Therefore, it can be said that they possess enough knowledge about the system and are qualified to provide meaningful and constructive feedback about the artefact, which is considered a part of the existing LA tool. The questions for the interview were formed according to the recommendations in (Bryman, 2016). The methodology is described in more detail in Chapter 4.

The interview was scheduled beforehand, at the date all the participants agreed on - 18th of May 2021. The chosen platform for the interview was a Voice Over Internet Protocol (VoIP) based program called “Zoom”, which became very popular during the COVID-19 world pandemic (Neate, 2020). This software solution allows recording the conversations, sharing the screen, sending files, and other useful features.

The interview included three main stages - introduction, main interview and outro. The details of each stage are presented below:

1. **Introduction.** During this stage, the contact between the parties of the interview was established, and several technical difficulties were identified and solved. After that, the interviewees were asked to give their consent regarding the recording of the conversation, which was subsequently provided, and the recording has started. After making sure that everything works, the screen was shared, and the interface of the developed artefact was demonstrated to the instructors. It then was briefly described, with the focus on main features and how they work. Both visualizations were shown, and their functionality and purpose were explained.
2. **Main interview.** After demonstrating the capabilities of the artefact, the prepared questions were asked in a sequence. Some of the questions required more details and explanations, which were also provided. Overall, this part can be described as a number of subsequent “question - answer” blocks, with additional clarifications where due.
3. **Outro.** When the list of the questions was exhausted, the parties of the interview engaged in a set of closing remarks and overall impressions regarding the interview process in

general and the artefact's functionality. Finally, the recording was stopped, and the interview was finished.

Initially, the interview itself was recorded, and then the feedback from the target group was transcribed into text, which is partly presented in the following sub-chapter.

7.2 Evaluation results

The evaluation focused on two main parts of the artefact - functionality and visualizations. Additionally, the feedback about more general topics has been collected, such as additions or changes to the existing features or visualizations, as well as suggestions and opinions about the artefact in general. The feedback about each of these three categories - functionality, visualizations and general, is described in more detail in the following sub-chapters.

7.2.1 Functionality

When answering the questions from this category, the respondents were positive about the functional parts of the artefact, like buttons and input fields, and their purpose. It was mentioned that it is easy to understand how the interactive elements work, what is the result of their operation, and what task they solve. In regard to the custom weights fields, the respondents acknowledged its potential, but not for their current course, since it has a relatively small number of students and such thorough customization of how each activity is weighed is not really that practical in the case of a small private online course.

The answers from the respondents regarding artefact's functionality are presented below:

- **Is it clear what is the purpose of the presented tool?**
  <u>Respondent 1</u>: *Yes.*

  <u>Respondent 2</u>: *Yes. The only thing I had to ask about is [the button] "Problems". But I think it's very intuitive when I think about it. So, I think that perhaps "Tasks" is more constructive and positive word than "Problems". I like that more.*

- **Was it easy for you to understand how the interactive elements of the interface work?**
  <u>Respondent 1</u>: *Yes, I think so.*

  <u>Respondent 2</u>: *Yes. I think the customization is really interesting.*

- **Would you use the custom-weighted grouping feature during the course?**
  <u>Respondent 1</u>: *I think the default view maybe is sufficient. It was fun to look at the other stuff, [...] but in the first expression it was easier to grasp maybe the default view I think.*

Respondent 2: *It's a very good question because for now we have done so small experiments, but I think in the future if we have larger groups we could experiment with… For example, if an exam is attached to videos, or to the assignments. I mean, it could be an interesting research question, if we could use it if we have larger groups. As we're doing now, it's not so immediate to see that, because we have so small numbers. But I think there's an interesting potential there.*

7.2.2 Visualization

The feedback about the visualization part of the artefact was also positive. The respondents mentioned that the default table view of activity and the Sankey chart are both easily recognizable, and their purpose and the information they convey are clear and understandable. The visualizations can be used to make course-related decisions, since it is easy to see the activity groups and each student in them. When comparing two types of the visualization, table and Sankey graph, both were accepted as useful representations of activity. However, the table view was chosen over the Sankey graph due to its simplicity and more intuitive look.

The answers from the respondents regarding the visualization part of the artefact are presented below:

- **Do the visualizations convey meaningful observations which you can use for making course-related decisions?**
  Respondent 1: *Yeah, I think so because it's easy to see who's watching videos, who's not watching the videos. And also [...] who's doing problems, who's not solving anything. So… it's an easy tool to see, both to get an input on what is used in the MOOC, and also on the individual level, what student is using what.*

  Respondent 2: *Yeah, these based on numbers, right? I think it's [the table view] is very intuitive, this one.*

- **Does the second visualization (Sankey chart) seem more informative to you than the table view?**
  Respondent 1: *I think the one that I'm looking at now [the default table view] is the easiest to look at I think. [after switching to the Sankey view] That one I think is more, it's more… You need to focus and think what does it really show me basically. So I thought maybe the first one [the default table view] is easy to look at.*

  Respondent 2: *I agree with the other respondent, because it's [the default table view] more immediate. But… I like both, ideally. But if I have to choose one that every teacher could easily use, I think that's that one [the default table view].*

- **Would you use the custom-weighted grouping feature during the course?**
  Respondent 1: *I think the default view maybe is sufficient. It was fun to look at the other stuff, [...] but in the first expression it was easier to grasp maybe the default view I think.*

7.2.3 General

In this part the respondents expressed several opinions about the possible changes and additions to the artefact:

- First suggestion was to expand the "Problems" category and separate it into several types of sub activities, such as quizzes, reflections and solving problems. It was mentioned that it will be quite useful to see the activity between these subcategories and to understand, which of them are most interacted with, so that it is possible to assess the necessity of their inclusion into the course program.
- Additionally, the naming of the button "Problems" was deemed to be not as clear and representative of what it actually stands for. The word "Tasks" was suggested instead, which should be more intuitive and clearer to the instructors.
- It was also mentioned that the weighting for the different activity types might not be as useful for smaller courses, such as the one that the respondents are a part of, but when the course has a relatively large number of students, forty or more, this feature might become quite useful, and one of the respondents expressed their interest in using it in one of the next courses.
- Both visualizations were described as meaningful, however, the table view was preferred by the respondents to be used in a course with relatively low number of students, while the Sankey graph is more fitting for the course with a high number of students. This is due to the ability of the Sankey graph to provide a quick overview of all the students in one visualization, while the table view requires manual scrolling through each category to find the particular student. However, this is more of a technical limitation, and both visualizations could be adjusted to be relatively equal in their perceived usability.
- Finally, the separation between the instructors and the students in the course activity grouping was suggested to be able to more easily distinguish the instructors from the students in both table and Sankey visualizations. Since the instructors might also be required to participate in the course, and possibly to complete the same tasks as the students, it would be useful to quickly see them in the list of students. This can be achieved by adding some sort of a special symbol next to their name or changing their colour in the activity group.

The answers from the respondents regarding the general questions about the artefact are presented below:

- **Which of the existing features would you like to be changed?**
  Respondent 1: *[...] If you could separate between quizzes and other tasks because when we're making the content it would be good to know if people are actually doing the quizzes, and are they doing the other stuff we add for them to do. So if it would be possible to separate them more - that would be good.*

  Respondent 2: *I think this is so much better than not having this, and I think it's so interesting. I'd like to have more experience using it. But I think that the "Problems" could be more nuanced, that would be great. It's hard for me to say anything more at this point of time, because we need more experience with it. And also it would be good to have a* *way to differentiate and take out the staff. Or just doing something to have it [the visualization] with staff and without staff. That would be great.*

7.2.4. Conclusion

Overall, the evaluation stage was very successful, and the feedback, albeit being very limited in the sample size, was enough to assess the first finished iteration of the artefact development process.

Next, the results of the whole research are discussed, alongside the methods and the conclusions for each stage.

Summary

This chapter overviewed the process and the results of the conducted evaluation of the implemented algorithm. The evaluation is an important step, which can either support or refute the current implementation of the algorithm. It is also possible during this step to discover the potential improvements and changes, which in-turn leads to understanding the direction for future work.

Chapter 8

Discussion

This chapter outlines the results of the evaluation and provides the answers to the proposed research questions. The discussion is divided into two parts - firstly, the results of the artefact development and how the methodology was applied for each stage, and secondly, the evaluation of the developed artefact, alongside the methodology, which has been used for this purpose.

8.1 Artefact development

The implementation of the artefact was executed according to the “Design science research” methodology. The methods and guidelines, which were provided by this methodology proved to be very helpful in conducting this research. Each stage of the development followed naturally one after another, carrying the nuances and observations to the following stages.

The methodology structure for this research, the “three cycle view”, which is presented in Chapter 4, can be detailed in the following form regarding the conducted research project:

- The environment for the study is the existing LA tool OXALIC together with a team, which manages this system. It provides a platform for the research and allows for feedback and corrections from the system owners during the development of the artefact.
- The knowledge base for the project is present as well in the form of a literature review, which makes it possible to understand the current state of the field, and to use the existing knowledge to formulate new goals for the research. The developed artefact should reach the set goals, thus contributing new insights to the knowledge base.

The principles of the chosen methodology are successfully applied in this research as well, and each of them can be described as follows:

1. The **“Design as artifact”** principle is satisfied, because during the research, the working artefact was created. The development started from understanding the status of the current scientific landscape in the chosen area - profiling of students based on their activity, which was presented in Chapter 2. After that, the methodology was selected and outlined in Chapter 4. Finally, the tools for development were identified, and the artefact implementation went through several stages according to the methodology. This process is described in Chapter 6.
2. The **“Problem relevance”** principle is satisfied, since the artefact solves a relevant problem, formulated after conducting the literature review, and supported by the organizational needs.

3. The **“Design evaluation”** principle is satisfied, due to the presence of the evaluation stage in the development cycle. The methodology for the evaluation was chosen and is outlined in Chapter 4. The evaluation itself, and its results are presented in Chapter 7.
4. The **“Research contribution”** principle is satisfied in a form of answering the specific research questions and by documenting the stages of the artefact development process.
5. The **“Research rigor”** principle is satisfied by using the existing knowledge and well-established techniques for creation of the artefact.
6. The **“Design as a research process”** principle is satisfied because the developed artefact operates in the existing organizational environment and helps solve the specific tasks.
7. The **“Communication of the research”**, since the results of the research are presented to both the professionals, who manage the existing system, as well as the users, that is the instructors, who operate the system. This is supported by being in constant communication with the owners of the platform, on which the artefact is based, and consulting with the main technician, who manages the current system.

The presented overview shows that the chosen methodology works well for this research. The methodology provides a good set of guidelines for the study and allows for a clear understanding of the steps that should be taken in order to achieve the proposed goals.

8.2 Evaluation

The format of the evaluation, which is outlined in Chapter 4, can be considered a good choice for this research project, as useful feedback was received from the respondents. The evaluation itself was conducted without noticeable problems, and the information, which was received from the respondents, provided insights on what should be adjusted and changed in the existing artefact. This allows transitioning to the next iteration of the development, during which the collected feedback can be applied, and the new, more refined features can be added as a result.

The feedback regarding the artefact was generally positive. Several suggestions were provided, which can potentially enhance the existing usability of the artefact. Both the functionality and the visualizations, provided by the artefact, were assessed positively. The respondents mentioned that they would use the features, which they were presented with, in their professional activities, and that it is clear, what is the purpose of the artefact and what meaningful information can be extracted from it.

8.3 Answering the research questions

The aim of this research has been to explore the ways of supporting instructors in assessing how active and engaged the students are in Open edX MOOCs. This was achieved by analysing the scientific literature regarding this topic, and by subsequently conceptualizing, implementing and evaluating the artefact, which aims to solve the aforementioned task.

The three research questions asked to guide the research were:

RQ1. How to identify engagement in Open edX MOOCs?

RQ2. What student profiles emerge through LA when it is applied to the activity data, and how can this be presented to instructors in the LA tool OXALIC?

RQ3. How do instructors use these student profiles to make course-related decisions?

8.3.1 RQ1. How to identify engagement in Open edX MOOCs?

The first research question was meant to uncover the ways of identifying and presenting the engagement of students with the help of tracking data, which is generated by the Open edX platform. In the foundation of this question lies the theory that since the actions of the students can be continuously tracked and stored during an online course, then it is possible to process this data and then form certain conclusions in accordance with the results of such processing. After conducting the research through the development of an artefact, which focuses on transforming the raw tracking data into easy-to-understand and intuitive bits of information, it can be concluded that the aforementioned theory is indeed correct. The documented process of creating the concept for the grouping algorithm, together with the description of the actual development and evaluation of the artefact, can be considered an answer to this research question. In order to identify engagement in MOOCs, which are based on the Open edX platform, one needs to (i) choose and implement the algorithm for processing the tracking data, including its transformation into a more convenient form to make further analysis easier and less time-consuming, and (ii) select and develop the visualizations for the processed information, so that it can be presented to the interested parties, such as instructors and researches, and the engagement of the students can be easily conveyed from these visualizations. Other, more specific steps can be mentioned too, such as organizing the storage for the tracking data, automating the process of saving, transformation and sending the tracking data inside the system and precomputing parts of the processed data to enhance the performance.

Ultimately, it can be concluded that in order to identify the engagement of the students in Open edX MOOCs, it is necessary to process the tracking data and the interaction events with the course, and then group the students into clusters, based on the amount of the interaction events, their frequency and other factors, such as types of the events and their priority in each given course.

It has also been discovered that even by identifying and forming a certain set of engagement groups in a particular course, the same groups can rarely be applied in the exact form to other courses. This is an issue that already exists in LA applications, called scaling learning analytics (Ferguson et al., 2014). This observation implies that each particular course can have different fundamental factors, which can influence how the engagement is identified and categorized. Therefore, it might be possible to create a specific and universal way, which establishes general rules about the engagement of the students, but in the end, the instructors themselves, with the help of the visualizations and additional features, such as adjusting the priority of each type of events, should decide, how to exactly outline and utilize the engagement of the course they are teaching.

8.3.2 RQ2. What student profiles emerge through LA when it is applied to the activity data, and how can this be presented to instructors in the LA tool OXALIC?

During the literature review, presented in Chapter 2, a multitude of student groups were identified across many MOOCs in a form of activity clusters. These clusters are heterogeneous in their nature, meaning that the groups of students, formed based on a specific dataset are not equal to the ones, which are formed using different dataset. This, and the fact that the size of the dataset influences the results of the clustering, suggested taking a different approach to solving the task of forming the engagement groups. As a result, a custom profiling algorithm was developed in the form of an artefact, which was specifically designed to be a part of the LA tool OXALIC for Open edX MOOCs. The following student profiles were formed in the process:

1. **No activity.** This profile includes students, who do not participate in the course in any way. This can be considered the most critical group, since students from this group require the most attention from the instructor, so that they can encourage the students to interact with the course.
2. **Low activity.** The students with a relatively low activity are placed in this group. These students rarely interact with the online course and may require additional encouragement from the instructor.
3. **Moderate activity.** This is the group, which consists of students with a relatively moderate engagement in the course material. Students from this group interact with the course and complete the tasks, but they can become less or more active with time, thus changing their current group to low or high activity.
4. **High activity.** Students with a relatively high engagement in the course are associated with this profile.
5. **Highest activity.** The students who interact with the course the most belong to this group. These are exceptional participants of the course who complete all the tasks and interact with the course the most.

The profiles are then presented to the instructors in the form of a dashboard in the LA tool OXALIC. These visualizations are based on the profiles, which are formed by processing the tracking data, generated by the Open edX MOOC. The instructors can interact with these visualizations and make certain course-related decisions based on the displayed information. This part is described in more detail in the next section.

8.3.3 RQ3. How do instructors use these student profiles to make course-related decisions?

After the student profiles are formed and presented to the instructors, they can be used to make course-related decisions. A few such actions are described below:

- **Change the weights of the activities.** By changing the values, which are used to form the activity scores of the students and assign them to one of the five profiles, the more important activities can be specifically targeted by the instructors in order to surgically identify the required groups of students. This allows instructors to dynamically prioritize, for example, video content over other types of activities in order to understand how engaged students are in the video-related material in the course.
- **Intervene when necessary.** The instructor can choose to act based on the visualizations, which represent the student engagement profiles in the course. For example, if the student is underperforming based on the tracking data, the instructor can then contact and communicate with them, providing support, encouragement and guidance to the next stage of the course the student should take. On the other hand, if the student shows increasing engagement, the instructor can encourage them to continue working with the course and try to achieve even better results.
- **Assess the state of the course.** The visualizations, which contain student engagement profiles, can be used by teachers to understand the overall state of the course, in which they are teaching. This makes it possible to quickly identify the students who underperform, which activities are most engaged in, and how students interact with each type of activity. Based on this the instructors can make certain course-related decisions, some of which are described in this part.
- **Plan changes and additions for existing and future courses.** By following the students' engagement throughout the course, the instructors can identify the parts, which have the lowest number of interactions from the students. This can indicate, among other things, that these parts are poorly realized and are not perceived by students as useful for them. Thus, based on this information, the least popular parts of the course can be adjusted and corrected, increasing the overall quality of the course and the amount of knowledge the students learn after finishing it.
- **Analyse the state of large courses.** MOOCs can contain hundreds of students and it is not feasible to try and understand how students behave by analysing the activities of each individual participant. By having the access to the visualizations, such as the ones in the developed artefact, the instructors can quickly understand, which students are, for example, underperforming, and act accordingly. Alternatively, groups of students, associated with the specific engagement profile, can be contacted at once by, for example, sending an email to the whole group, providing guidance or updates on their performance in the course.

Additionally, the target group of instructors positively assessed the presented artefact during the evaluation, which indicates that the implemented functionally and visualizations can indeed be used in the context of LA tool OXALIC and Open edX MOOCs to make course-related decisions.

Summary

This chapter presented the author's point of view on the conducted research, its results and how the study answers the research questions. It is important to interpret the knowledge gained by this study in the form of the answers to the research questions, so that the contribution to the knowledge base is clear.

Chapter 9

Conclusion and future work

This chapter concludes the conducted research and provides thoughts about possible directions for future opportunities regarding the theme of this research. Several limitations, which were encountered in the process of this research, are presented as well.

9.1 Conclusion

The main goal of this thesis was to discover how LA can be effectively and efficiently applied in the context of Open edX MOOCs and LA tool OXALIC, and how it can benefit the stakeholders, such as instructors and researchers, in understanding the behaviour of the students participating in the course and allowing the instructors to make course-related decisions. To achieve this goal, the literature was thoroughly reviewed, and several potential research areas were uncovered. One of them, namely clustering, or grouping, of students based on their activity was chosen as the main focus for the thesis. The methodology then was established and outlined for both the development of the profiling artefact, as well as for the process of its evaluation. The resulting artefact was successfully developed and evaluated with the help of the chosen methodology, which proved to be very useful and feasible for the purpose of the research. The final observations show that the artefact was able to demonstrate how LA can be applied to the tracking data, generated by the Open edX platform, and what exactly is required to achieve this outcome. The resulting visualizations were met with positive feedback, which indicates that the research, albeit with several limitations, which influenced the whole study, was moving into the right direction. The documented process of the development, together with the evaluation, is considered as the main contribution of this thesis, since this information can be used in development of a more complex system, or to continue building new features on top of the existing functionality. Several opportunities regarding the additional applications of LA were discovered and are presented in the following sub-chapter.

To summarize, this research can be considered as overall successful. It followed a well-defined methodology and showed promising results, which were accepted by a rather limited, but nonetheless meaningful number of actual professionals, who expressed their positive opinions regarding the application of LA in Open edX MOOCs in a form of the artefact, resulting from this study.

9.2 Limitations

The following limitations were encountered during the writing of this thesis:

- **Global pandemic and lockdowns.** Particularly, the complications caused by the strict restrictions and lockdowns posed by the government and the authorities. Before they were applied, the research went very well due to the ease of communication, and availability of other people, who could contribute to the process of the research, such as fellow students. After the establishment of the lockdowns and other restrictions, the flow of work was disrupted, and slowed down considerably. This resulted in less features being implemented than planned.
- **Lack of data to work with.** Originally, the amount of data available for the research was very limited, which caused the research to stall and compromise on its original goals. Instead of having tracking data from courses with a substantial number of students, only the data from short courses with a small number of students was available, which made it impractical to apply certain LA methods, like Machine Learning, which could arguably produce better results than the proposed profiling algorithm.
- **Difficulties with obtaining the data.** The process of obtaining the data has also influenced the pace of the research, since it required a lot of time to receive the data from the owners of the courses. This is mainly due to the fact that this particular data contains some pieces of personal information, which can potentially be used to identify the people who participated in the course, which requires additional security steps to be taken before the data can be shared. This process was handled on the side of the system owners.
- **Small sample size during evaluation.** The results of the evaluation might be considered quite limited, because only two instructors participated in it. This might have reduced the amount of feedback about the artefact and its functionality, which can be considered as one of the main contributions of this research, therefore making it less useful to other researchers.

9.3 Future work

The conducted research naturally cannot in any way, shape or form be considered as finished, since the original plan has not been fully realized during the period of the research due to the aforementioned limitations. However, several opportunities for future work were uncovered during the study. The unrealized and potential additions to the existing artefact are presented below:

- **Clusters with typical activity.** By analysing the tracking data from multiple past courses, a universal set of typical clusters can potentially be outlined. For example, a specific course can have three typical clusters of students based on the contents of the course. The outliers, which do not fit into these typical clusters, should be then inspected by the instructor, since they are acting differently from other students and may require additional attention.
- **Learning sessions.** The concept of a learning session can be also researched to provide an extended overview of the student’s activity in the course. A learning session can be

described as a period of time when the student is interacting with the system, somewhat reminiscent of the seminar session in the university. To achieve that, the way of identifying the start and the end of the session should be clearly established, as well as the method of selecting the appropriate amount of time between the events. The resulting information can also be used as an addition to understand the engagement of the students in the course.

- **Additional visualizations.** The existing visualizations can also be expanded, and novel ways of displaying the processed tracking data can be conceptualized. As an example, a representation of each student's activity in the form of a web graph can be added. This way, the instructor will have a clear grasp over how the student spends time with the course. More novel visualization approaches, such as the one mentioned in Chapter 5, can be prototyped as well. However, this will require an extended evaluation process, since novel approaches to visualization can be less desirable by the users of the system compared to the well-established ones, such as Sankey chart or a web graph.
- **Interventions.** Another potential area of research are interventions from the instructors during the course based on the information, which is provided by the application of LA techniques. Intervention is a certain action from an instructor, which is aimed at providing directions for underperforming students, or appreciating the students with good results in the course. In the case of this thesis, the interventions based on engagement can be looked into in more detail. The questions such as “Is it possible to make interventions based on the engagement groups?” and “How effective is this kind of intervention?” could be answered. For example, if students are categorized as ones with a low activity, and the instructor sends them an encouraging message using the interface of the artefact, will these students’ performance in this course increase? If so, such functionality could be considered very useful for instructors, especially in larger courses, since it lessens the burden of managing every single student and contacting them one-by-one.
- **Flexible time periods for grouping.** Finally, the ability to dynamically change the periods of time, on which the engagement is based, can be explored. Nuances such as bursts of activity in certain periods, and their influence on the final results of each student can potentially uncover the more productive ways to study in an online course environment, since it is different from traditional ways of studying, such as seminars.

Overall, the conducted research shows great potential, and additional research in this area can uncover interesting and useful contributions, which will benefit the whole educational sector.

References

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics, 37,* 13-49.

van den Beemt, A., Buijs, J., & van der Aalst, W. (2018). Analysing structured learning behaviour in massive open online courses (MOOCs): an approach based on process mining and clustering. *International Review of Research in Open and Distributed Learning, 19*(5).

Bozkurt, A., Keskin, N. O., & De Waard, I. (2016). Research trends in massive open online course (MOOC) theses and dissertations: Surfing the tsunami wave. *Open Praxis, 8*(3), 203-221.

Brinton, C. G., Buccapatnam, S., Chiang, M., & Poor, H. V. (2016). Mining MOOC clickstreams: Video-watching behavior vs. in-video quiz performance. *IEEE Transactions on Signal Processing, 64*(14), 3677-3692.

Bryman, A. (2016). *Social research methods*. Oxford university press.

DjangoCon US. (2018). *DjangoCon US 2018 - Anatomy of Open edX - a modern online learning platforms... by Nate Aune* [Video]. YouTube. https://youtu.be/7BBvxn4o_uE.

Dresch, A., Lacerda, D. P., & Antunes, J. A. V. (2015). Design science research. In *Design science research* (pp. 67-102). Springer, Cham.

Ferguson, R., Clow, D., Beale, R., Cooper, A. J., Morris, N., Bayne, S., & Woodgate, A. (2015). Moving through MOOCS: Pedagogy, learning design and patterns of engagement. In *Design for teaching and learning in a networked world* (pp. 70-84). Springer, Cham.

Ferguson, R., & Clow, D. (2015). Examining engagement: analysing learner subpopulations in massive open online courses (MOOCs). In *Proceedings of the fifth international conference on learning analytics and knowledge* (pp. 51-58).

Ferguson, R., Clow, D., Macfadyen, L., Essa, A., Dawson, S., & Alexander, S. (2014). Setting learning analytics in context: Overcoming the barriers to large-scale adoption. In *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 251-253).

Json.org. (2002). *Introducing JSON*. Retrieved from https://www.json.org/json-en.html.

Khalil, M., & Belokrys, G. (2020). OXALIC: an Open edX Advanced Learning Analytics Tool. In *2020 IEEE Learning With MOOCS (LWMOOCS) (pp. 185-190)*. IEEE.

Khalil, M., & Ebner, M. (2015). Learning analytics: principles and constraints. In *EdMedia+ Innovate Learning* (pp. 1789-1799). Association for the Advancement of Computing in Education (AACE).

Khalil, M., & Ebner, M. (2016). What massive open online course (MOOC) stakeholders can learn from learning analytics?. *arXiv preprint arXiv:1606.02911*.

Khalil, M., & Ebner, M. (2017). Clustering patterns of engagement in Massive Open Online Courses (MOOCs): the use of learning analytics to reveal student categories. *Journal of computing in higher education, 29*(1), 114-132.

Khalil, M., Kastl, C., & Ebner, M. (2016). Portraying MOOCs learners: A clustering experience using learning analytics. *Proceedings of the European Stakeholder Summit on experiences and best practices in and around MOOCs (EMOOCS 2016)*, 265-278.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 170-179).

Lau, K. V., Farooque, P., Leydon, G., Schwartz, M. L., Sadler, R. M., & Moeller, J. J. (2018). Using learning analytics to evaluate a video-based lecture series. *Medical teacher, 40*(1), 91-98.

Len-Urritia, M., Cobos, R., & Dickens, K. (2018). MOOCs and their influence on higher education institutions: Perspectives from the insiders. *Journal of New Approaches in Educational Research (NAER Journal), 7*(1), 40-45.

March, S. T., & Storey, V. C. (2008). Design science in the information systems discipline: an introduction to the special issue on design science research. *MIS quarterly*, 725-730.

Mozilla. (2021). *About JavaScript*. Retrieved from https://developer.mozilla.org/en-US/docs/Web/JavaScript/About_JavaScript.

Neate, R. (2020). Zoom booms as demand for video-conferencing tech grows. *The Guardian, 31*.

Ruipérez-Valiente, J. A., Munoz-Merino, P. J., Gascon-Pinedo, J. A., & Kloos, C. D. (2016). Scaling to massiveness with analyse: A learning analytics tool for open edx. *IEEE Transactions on Human-Machine Systems, 47*(6), 909-914.

Schumacher, C., & Ifenthaler, D. (2018). Features students really expect from learning analytics. *Computers in human behavior, 78,* 397-407.

Shukor, N. A., & Abdullah, Z. (2019). Using learning analytics to improve MOOC instructional design. *International Journal of Emerging Technologies in Learning (iJET), 14*(24), 6-17.

Stanford News. (2013). *Stanford to collaborate with edX to develop a free, open source online learning platform.* Retrieved from https://news.stanford.edu/news/2013/april/edx-collaborate-platform-030313.html.

The State of JavaScript. (2021). *Front-end Frameworks.* Retrieved from https://2020.stateofjs.com/en-US/technologies/front-end-frameworks/.

Torre, M. V., Tan, E., & Hauff, C. (2020). edX log data analysis made easy: introducing ELAT: An open-source, privacy-aware and browser-based edX log data analysis tool. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 502-511).

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior, 89*, 98-110.

Vuetify. (2021). *Introduction*. Retrieved from https://vuetifyjs.com/en/introduction/why-vuetify/.

Zhang, G., Zhang, Y., & Ran, J. (2018). Research on Clustering Mining and Feature Analysis of Online Learning Behavioral Data Based on SPOC. In *2018 13th International Conference on Computer Science & Education (ICCSE)* (pp. 1-6). IEEE.

Zorrilla, M., & de Lima Silva, M. (2019). Sociograms: An Effective Tool For Decision Making in Social Learning. *Technology, Knowledge and Learning, 24*(4), 659-681.

Appendix A

A set of questions for the evaluation interview

Functionality:

- Is it clear what is the purpose of the presented tool?
- Was it easy for you to understand how the interactive elements of the interface work?
- Are the interactive elements like buttons and input fields intuitive to use?
- Would you use the custom-weighted grouping feature during the course?

Visualization:

- Do the visualizations convey meaningful observations which you can use for making course-related decisions?
- Are the presented visualizations appropriate for the data they try to display?
- Are the presented visualizations simple enough to understand their purpose?
- Would you use this tool in your professional work to make course-related decisions? Could you elaborate more on why it seems useful or not useful to you?
- Does the second visualization (Sankey chart) seem more informative to you than the table view?

General:

- Which of the existing features would you like to be changed?
- Which additional features would you prefer to be added to the current state of the engagement tool?
- Do you have any other feedback about the engagement tool or how it works in general?