Visualizing Content Exploration Traces of MOOC Students

Tobias Rohloff, Max Bothe, Christoph Meinel
Hasso Plattner Institute, Potsdam, Germany
{tobias.rohloff,max.bothe,christoph.meinel}@hpi.de

ABSTRACT: This workshop paper introduces a novel approach to visualize content exploration traces of students who navigate through the learning material of Massive Open Online Courses (MOOCs). This can help teachers to identify trends and anomalies in their provided learning material in order to improve the learning experience. The difficulty lies in the complexity of data: MOOCs are structured into multiple sections consisting of different learning items and students can navigate freely between them. Therefore, it is challenging to find a meaningful and comprehensible visualization that provides a complete overview for teachers. We utilized a Sankey diagram which shows the students' transitions between course sections by grouping them into different buckets, based on the percentage of visited items in the corresponding section. Three preceding data processing steps are explained as well as the data visualization with an example course. This is followed by pedagogical considerations how MOOC teachers can utilize and interpret the visualization, to gain meaningful insights and execute informed actions. At last, an evaluation concept is outlined.

Keywords: MOOCs, Learning Analytics, Content Exploration, Sankey Diagram

1 INTRODUCTION

Massive Open Online Courses (MOOCs) are attended by thousands of learners (Shah, 2018), which makes it hard for teachers to keep the overview of their students' progress. Dashboards have been proven to be a helpful tool by providing different data visualizations and statistics (Klerkx, 2017), as implemented by many MOOC platforms. Some of these visualizations are easy and intuitive to understand and some require a slightly longer learning curve. Especially when the data becomes more complex, it gets harder to understand the visualizations. One difficult case in MOOCs is to comprehend how learners navigate through the course material. Courses are usually structured into sections, which represent course weeks or topic chunks. Each section consists of different learning items, mostly video lectures, texts, exercises and quizzes. Even if an order is given through the structured material, learners can explore the course in any sequence or skip content at all. Thus, it is complicated to visualize these content exploration traces for all students of a course. However, teachers could benefit for example by identifying anomalies within their provided material.

With this workshop paper, we introduce one possible visualization technique that we implemented for the HPI MOOC platform[1]. We looked at approaches from other disciplines, like conversion rates in web analytics (Zheng, 2015) and funnel charts. In the end, we decided to implement a Sankey diagram which displays transitions between course sections based on their amount of visited learning items of each student. This paper explains how the data is being processed and visualized. Additionally, it

[1] https://open.hpi.de/

discusses how MOOC teachers and instructors can utilize and interpret the visualized data, to obtain meaningful insights from their students' learning behavior and execute informed actions and interventions. At last, an evaluation concept is presented to investigate the helpfulness and comprehensibility of the visualization, amongst other things.

2 DATA PROCESSING

In order to visualize the student's content exploration traces with a Sankey diagram, the captured interaction data needs to be processed first. The platform stores all user interaction events in redundant analytics storages within a central learning analytics service (Rohloff, 2018). These storages are realized with different database technologies to enable different query techniques (like SQL or NoSQL) for an optimized performance of each implemented metric. The events are structured in an xAPI-alike format (Renz, 2016). For the intended visualization, the users' learning item visits need to be processed into aggregated nodes and links for the Sankey diagram. Therefore, a new metric was introduced within the learning analytics service, which takes care of three processing steps (one database query and two post-processing steps) to compute the final data structure. These three processing steps are explained briefly in the following sections. The data processing performance depends on the size of the data basis. With our current infrastructure and more than 360,000,000 user interaction events today on a single platform, the load is too high to generate the data on-demand with every request. Therefore, we decided to process the data once per day for each course and store the results. The persisted data is then displayed to teachers.

2.1 Process Raw Events into Visit Counts per Section for each User

The first step processes the raw events stored in the database and calculates the unique item count per section for every user. Therefore, the SQL-based storage was used (PostgreSQL). The data is stored in an event table and queried. It returns a list of dictionaries with each element containing a unique combination of a `user_id` and `section_id`, as well as the distinct visited `item_count` (visit count). Additionally, only events captured during the regular course runtime have been taken, since self-paced course activity should be examined separately. The maximum length of this list is $u * s$, where u is the number of users and s is the number of sections.

2.2 Process Visited Percentage for all Sections for each User

Now, for each user, the visited percentage for all sections of a course is calculated based on the visit count from the previous step. This ensures that values for all sections are generated, even for sections without any visits. Additionally, the visited percentages are sorted according to the section positions. This results in a dictionary where every `user_id` points to a sorted list of visited percentages, representing the different course sections.

2.3 Process Bucketized Nodes and Node Links

In the third step, the different percentage values of every user are aggregated. Each Sankey node layer will represent a different course section. The nodes of one layer will display different visited percentage buckets. Therefore, the visited percentages ranging from 0.0 to 1.0 were split into specific intervals. We decided on the following configuration:

$$[0.0]\ \cup\ ]0.0, 0.2[\ \cup\ [0.2, 0.4[\ \cup\ [0.4, 0.6[\ \cup\ [0.6, 0.8[\ \cup\ [0.8, 1.0[\ \cup\ [1.0]$$

We treated no visits (0.0) and all items visited (1.0) as special cases to identify learners who never showed up in a section (no-shows) and very engaged learners (completers). Now, we had to determine for each possible link between the nodes of adjacent layers (source node of layer a to target node of layer b) the number of users. This resulted in $i * i * (s - 1)$ links, where i is the number of defined intervals and s is the number of sections. For our configuration of 7 intervals and a course with 7 sections, this would produce 294 links for all nodes. The size of each node can be derived from the links, by summing up the corresponding user counts. The number of nodes is $i * s$, resulting in 49 nodes for our given example. With all nodes and links in place, the Sankey diagram can be drawn.

3 DATA VISUALIZATION

To render the diagram as part of the platform's web-based teacher dashboard a D3[2] Sankey plugin[3] was used. The output for our example configuration is shown in Figure 1. The 7 vertical node layers represent the course sections (the section labels are omitted in the figure). Each layer consists of 7 nodes, which are annotated with the corresponding visited percentage intervals. The nodes and links are color-coded, ranging from red for no visited items to green for all items visited. The colors of the inner nodes are interpolated with a ratio based on the nodes' interval threshold. This enables a dynamic colorization based on the interval configuration. The links have the same color as their target node, with a slight transparency to display overlaps.

The specific user count value of a node or link is shown when hovered in the web browser. Additionally, all links connected to a node are highlighted when hovering the node. This can help teachers to comprehend cohorts of students with unusual behavior. For example, if a larger group of students, who visited a lot of learning items in a section, only visits a few items of the following section.

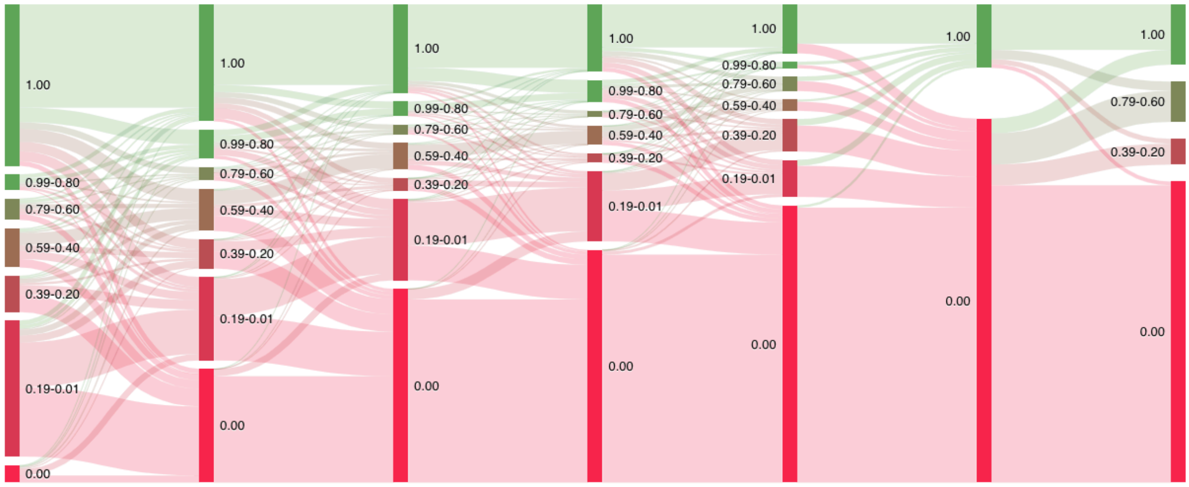

Figure 1: Content Exploration Traces of a MOOC with 7 Sections visualized as a Sankey Diagram.

[2] https://d3js.org/

[3] https://github.com/q-m/d3.chart.sankey

The course chosen to visualize Figure 1 included 5,284 students. The first five sections were successive course weeks, each consisting of video lectures, self-tests and a weekly graded quiz. In the sixth section a final exam was conducted, followed by an "I like, I wish" final section to gather the students' feedback. It can be seen that in this example, that the majority of completers also complete the following section and only a few of them visit fewer items. Also, only a minority of no-shows come back to visit content in the next sections. The midfield shows a more diverse picture and there is also a general trend visible, that the number of no-shows increases from section to section, who will end as drop-outs most probably. However, the reasoning behind requires the knowledge of a human expert. Visualizations like this can only support evaluations and decisions resulting from it.

4 PEDAGOGICAL VALUE

After the presentation of technical implementation details, this section will introduce how teachers and instructors in MOOCs can utilize this diagram and benefit from it. Related and similar visualizations either used stacked bar charts for a weekly student participation overview, or state transition diagrams for learning items (Coffrin, 2014). However, our proposed diagram provides the advantages of both approaches: it shows a complete course overview with different stacked user subgroups and also displays the transitions of these subgroups between course sections. This should serve as a starting point for teachers to get a first overview of student activity and their content exploration traces in a course. Thereby, the diagram is meant to complement other visualizations which focus on more detailed aspects of a course, like assignment grade distributions, video navigation charts or forum activity graphs (Stephens-Martinez, 2014).

Therefore, it is placed in the teacher's central course dashboard as one of the first visualizations. Based on the displayed data, the teacher is able to quickly spot unusual behavior and anomalies across the whole course. Then, the corresponding content can be delimited and identified to further examine the issue and execute informed interventions. Thereby, the two main elements of the Sankey diagram can be used to interpret the data. First, the stacked nodes show how many active and less active students participate in a certain course section. These numbers can be either compared with other course sections, other iterations of the same course, or different courses to see how well perceived a certain section is. Second, the links show the transitions of different engaged student subgroups, which can be helpful in various ways. No-shows in one section are highly likely to be no-shows in the following section as well since they never appear in the course again and can be ignored. However, a transition of a large portion of highly engaged users in one week to a low activity group in the next week is unusual and may indicate an issue in the preceding week, like too tiring video lectures or a too difficult weekly exam. But here the interpretation possibilities are reaching the limits of this diagram and depend on the respective case. Nevertheless, further investigations can be done with other visualizations which focus on certain aspects instead of a complete overview, as discussed before.

A future evaluation is necessary to test our assumptions if the Sankey diagram is interpretable enough for real-world MOOC instructors and if they consider it as helpful to monitor their students' activity to make informed and meaningful actions. This evaluation will be done on different deployments of the HPI MOOC platform. It is planned to do this separately, but also as part of a larger teacher dashboard evaluation. Interviews can be used for qualitative analysis, and for quantitative analysis surveys and usage data. The usability and comprehensibility will be investigated, but also the specific value as a learning analytics tool, e.g. by measuring its EFLA score (Scheffel, 2017). Even if the diagram is able to realize the goal of an overview of a whole course, teachers need to explore the details of an identified trend or anomaly to better examine the cause. Therefore, the diagram must be complemented with more detailed visualizations, which show what happens inside sections between different learning items. Here, also the difficulty of quizzes or comprehensibility of videos can be utilized for example, next to the item visits. This needs to be implemented as well before the evaluation is conducted.

5 CONCLUSION

This paper introduced a novel approach on how a Sankey diagram can be used to visualize students' content exploration traces between sections of a MOOC. Based on captured user interaction events, three processing steps were explained to generate the data for the nodes and links of the Sankey diagram, by using vertical node layers as a representation of different MOOC sections. Each node displays the share of a certain interval of the total visited learning items percentage of a section. By interacting with the Sankey diagram, the teacher can highlight connected notes to comprehend cohorts of students with unusual behavior. An example is shown for a real-world MOOC with possible conclusions. Additionally, the pedagogical value of the visualization was discussed, how instructors can use and interpret the visualization and gain meaningful insights to take informed actions. Also, an evaluation concept was outlined to test our assumptions and examine the helpfulness and comprehensibility. All in all, this work showed a possibility to display a complete overview for MOOC teachers, how their thousands of students navigate through the course material.

REFERENCES

Coffrin, C., Corrin, L., de Barba, P., Kennedy, G. (2014, March). *Visualizing patterns of student engagement and performance in MOOCs*. Paper presented at the 4th International Conference on Learning Analytics and Knowledge. https://doi.org/10.1145/2567574.2567586

Klerkx, J., Verbert, K., Duval, E. (2017). Learning Analytics Dashboards. In Lang, C., Siemens, G., Wise, A., Gasevic, D. (Eds.), *Handbook of Learning Analytics* (Chapt. 12). Society for Learning Analytics Research.

Renz, J., Navarro-Suarez, G., Sathi, R., Staubitz, T., & Meinel, C. (2016, April). *Enabling Schema Agnostic Learning Analytics in a Service-Oriented MOOC Platform*. Paper presented at the 3rd ACM Conference on Learning @ Scale. https://doi.org/10.1145/2876034.2893389

Rohloff, T., Bothe, M., Renz, J., & Meinel, C. (2018, September). *Towards a Better Understanding of Mobile Learning in MOOCs*. Paper presented at the 5th IEEE Conference on Learning with MOOCs. https://doi.org/10.1109/LWMOOCS.2018.8534685

Scheffel, M., Drachsler, H., Toisoul, C., Ternier, S., & Specht, M. (2017, September). *The Proof of the Pudding: Examining Validity and Reliability of the Evaluation Framework for Learning Analytics*. Paper presented at the 12th European Conference on Technology Enhanced Learning. https://doi.org/10.1007/978-3-319-66610-5_15

Shah, D. (2018, January). By The Numbers: MOOCS in 2017 [Blog post]. Retrieved from https://www.class-central.com/report/mooc-stats-2017/

Stephens-Martinez, K., Hearst, M. A., Fox, A. (2014, March). *Monitoring MOOCs: Which Information Sources Do Instructors Value?* Paper presented at the 1st ACM Conference on Learning @ Scale. https://doi.org/10.1145/2556325.2566246

Zheng, J., Peltsverger, S. (2015, January). Web Analytics Overview. In Khosrow-Pour, M. (Ed.), *Encyclopedia of Information Science and Technology, Third Edition* (Chapt. 756). IGI Global.