MOOC Big Data: A win-win situation

Karan Shah, College of Computing, Georgia Institute of Technology

I. Abstract

With the exponential rise in the number of MOOCs, a substantial amount of data has been and is being generated by these courses. This data can benefit everybody, both the producers and consumers. This paper explores some ways in which this data can be used.

II. Background

Massive Open Online Courses (MOOCs) are online courses usually offered by universities and other education providers on a large scale. Right now, there is are a lot of high quality MOOCs available for free from top institutions such as MIT, Georgia Tech, Harvard etc. While free high quality educational content is truly a great story, there is a much more significant impact that is just under the surface in MOOC data, and may prove to be far more significant in the long run. The ability for MOOC providers to generate massive amounts of useful data from course enrollments ranging from 10,000 to 100,000 students will enable providers to not only lower the cost of education, but also improve outcomes.

III. Uses of MOOC big data Via **optimization, experimentation and personalization**, MOOCs have an opportunity to dramatically improve the quality of education in a way that has not been possible up until now.

Optimization

Like most tech startups, MOOCs like Coursera collect data for every action (or inaction) performed by a student – when a student pauses a video, increases playback speed, answers a quiz question, revises an assignment, or comments in a forum. This microscopic level of data, when collected at the scale that MOOCs operate on, is perhaps the most valuable in identifying the root cause of failure by students. This is helpful in finding fundamental errors in the way of teaching. If the majority of students in a cohort struggle with a particular topic or problem, it means that the instructor must figure out a way to make the content more manageable. The widespread reach of MOOCs makes it possible to optimize courses by recognizing areas where content is insufficient to deliver on learning goals – something that would be impossible to identify in a university classroom of 200 students. If a test question is answered incorrectly, or if students lose focus during a specific point in the course, data can direct the course creators to go back into the curriculum to add or modify the content to better suit the learning outcome. More than anything, data and scale will enable instructors to have actionable feedback about their courses.

Using the data from the online version of Fall 2012 offering of 6.002x: Circuits and Electronics, researchers at MIT have built a system that predicts stopout rate in MOOCs quite accurately.[1]

The raw data that the team used included 154,763 registered learners, 17.8 million submission events, 132.3 million navigation events and about 90,000 forum posts. They analyzed this data based on 18 parameters. By analyzing this data, they not only proved that stopout prediction is a tractable problem, they did it with accuracy. The accuracy rate was as high as 95% when the prediction was for 1 week in the future and 70% for predictions for the end of the course. Another innovation in this study was that they asked students about what parameters to use to optimize the course. So some of the 28 parameters were crowdsourced.

The same team published another paper about their analytics platform which helps analyze student behavior during problem solving in MOOCs [2]. Using the data from the above mentioned 6.002x course, they compared how students from different regions and cohorts compare to each other. Using their analysis, instructors can learn whether there is a significant difference in these variables: average response formulation duration, average problem attempts and average resource consultations for labs vs. homeworks.

Experimentation

Because most of the MOOC aggregators are tech startups, they utilize A/B testing. A/B testing means comparing multiple versions of content to determine which one is most consumed. A/B testing was impossible in traditional classrooms because of the small number of students which doesn't give a statistically representative sample. However, it is possible in MOOCs which operate at a much larger scale.

Using such experimentation, MOOCs can continuously improve courses, but more importantly, they can improve the entire educational process, and allow data rather than intuition drive results. Whereas most of the classes in colleges likely change very slightly from semester to semester, this new data will allow MOOC professors and administrators to evolve their classes on a weekly basis. Although a typical A/B test focuses on a very narrow hypothesis (i.e. blue button vs. green button), over time hundreds of such tests could eventually evolve the very nature of online courses themselves. Another aspect of experimentation is the peer review process. When using peer review, the outcomes of a course might change from cohort to cohort. The way a cohort thinks collectively is likely to change from time to time because of events in the real world. Using feedback from one cohort to tweak the course for the next cohort will also increase the quality of the courses.

At the Learning@Scale Conference 2014 that took place here in Atlanta, a team from UIUC presented a paper that showed that the active involvement of the professor did not affect student performance in their MOOC[3]. They did this by using A/B testing. An A/B test was used to randomly assign MOOC participants in either a control group (with no instructional interaction) or an intervention group (in which the professor and teaching assistants responded to comments in the discussion and complied summary weekly feedback statements) to identify the differences in learning outcomes, participation rates, and student satisfaction. The course was "Introduction to Sustainability", delivered on Coursera by UIUC. They noted that the difference in performance in both the groups was statistically insignificant. Such studies help instructors manage their time and energy better and let them focus on better content.

Personalization

Websites and apps spend a lot of time, effort and resources to personalize their services. This is because personalization leads to a better user experience and enables the service to have a greater impact on the user. It's the same with education. Online courses driven by software make it possible to remove the "one size fits all" model and instead deliver content to students in the format and method that most suits the way that they learn. This type of personalization is likely some time away, but it's easy to envision how a user's past learning experiences could be used to improve the way the user is taught in the future. This becomes even more significant when we consider online learning becoming a part of the K – 12 education system, and being able to build a learning profile that takes into account many years of a user's life. In this new model, content and delivery can be better customized for each student, so if one student required visual aids to learn, whereas another student prefers to read, and another student requires an assignment, each will get what they need in order to learn and succeed. Companies like Coursera are already attempting this, with services like weekly emails that contain links to courses that the user might like. Such recommendations are generated by utilizing user data. A lot of courses, especially programming courses, give very detailed and personalized outputs on a large scale without the need for human intervention.

Another application of this content recommendation system is that by analyzing how students fare in the courses they are currently taking, recommendations can be personalized based on skill levels such that slow learners get reinforced with appropriate learning content which will also help them in the current course. A team from VIT University, India has implemented this in an edX course using an algorithm called CBRL which is showing good results in limited student sets[4].

IV. Conclusion

As explained above, big data will lead to better quality content for consumers (students) and feedback producers (universities and companies). Utilizing these vast troves of data will be pivotal in giving MOOCs credibility and making them a part of the standard education systems.

Bibliography

[1]: Colin Taylor, Kalyan Veeramachaneni, Una-May O'Reilly, “Likely to stop? Predicting Stopout in Massive Open Online Courses”
arXiv:1408.3382v1

[2]: Fang Han, Kalyan Veeramachaneni and Una-May O'Reilly, “Analyzing Millions of Submissions to Help MOOC instructors Understand Problem Solving”, DDE@NIPS 2013: Data Directed Education.
URL: http://groups.csail.mit.edu/EVO-DesignOpt/groupWebSite/uploads/Site/ProblemAnalytics.pdf

[3]: Jonathan H. Tomkin, Donna Charlevoix, “Do professors matter?: using an a/b test to evaluate the impact of instructor involvement on MOOC student outcomes, Proceedings of the first ACM conference on Learning @ scale conference, March 04-05, 2014, Atlanta, Georgia, USA” [doi>10.1145/2556325.2566245]
URL: http://doi.acm.org/10.1145/2556325.2566245

[4]: Raghuveer, V.R.; Tripathy, B.K.; Singh, T.; Khanna, S., "Reinforcement learning approach towards effective content recommendation in MOOC environments," in*MOOC, Innovation and Technology in Education (MITE), 2014 IEEE International Conference on* , vol., no., pp.285-289, 19-20 Dec. 2014
doi: 10.1109/MITE.2014.7020289
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7020289&isnumber=7020228