# How Do Search Engines Work? A Massive Open Online Course with 4000 Participants

Ralf Krestel and Julian Risch

Hasso-Plattner-Institut, Prof.-Dr.-Helmert-Str. 2–3, 14482 Potsdam, Germany
firstname.lastname@hpi.de

**Abstract.** Massive Open Online Courses (MOOCs) have introduced a new form of education. With thousands of participants per course, lecturers are confronted with new challenges in the teaching process. In this paper, we describe how we conducted an introductory information retrieval course for participants from all ages and educational backgrounds. We analyze different course phases and compare our experiences with regular on-site information retrieval courses at university.

**Keywords:** Massive open online course, MOOC, search engines, IR teaching

## 1 Information Retrieval for the Masses

Massive open online courses (MOOCs) are established as a way of flexible learning with millions of participants worldwide. While the gold rush five years ago has lead to proclamations that MOOCs are the future of education [29], today, with a more mature field and years of experience with MOOC platforms, the goal and vision of MOOCs has shifted [35]. Instead of offering a complete curriculum, individual MOOCs are focusing on introductory lectures or specific technologies. Especially computer science topics are offered by various commercial platforms and universities. These courses are typically targeted towards (computer science) students and people working in the IT industry. These groups are used to online tutorials, e-learning environments, and flexible learning schedules. Also, they use MOOCs as an alternative to traditional courses at their university or off-the-job training to which most of them could get access to. But, the penetration of everyday life with Information, especially by the World Wide Web, necessitates a confident handling of new technologies not only by expert users. Search engines and information retrieval concepts in general are heavily used by people of all ages and educational backgrounds. While some high schools promote digital literacy, most people, especially older citizens, have never learned how the underlying technology they are using everyday works. This gap between high school graduates and IT experts can be filled by MOOCs as an alternative to adult education centers. Also, in countries where education is very expensive or not accessible or available, MOOCs provide a means of inexpensive teaching [8, 13].

Information retrieval is not the only computer science topic relevant to a broader audience, but with search engines there exists a point of interaction that

everybody knows and uses daily. Therefore are information retrieval concepts and understanding how web search engines work important components for digital literacy in today's societies. MOOCs on these computer science topics can be a vehicle to teach and inform all living in modern societies.

In contrast to traditional classroom teaching, a set of specific challenges needs to be addressed when teaching MOOCs. The most important one is the heterogeneity of the participants. Various age groups, cultural and educational backgrounds, and also motivation of the individuals differ largely. On the other hand, the technology of MOOC platforms allows usually to adjust the speed of learning individually and thus the trade-off between overburdening beginners and not challenging more experienced participants is delegated back to the user. In practice, this means recorded video lectures can be skipped or played back at a higher speed if the user is already familiar with the presented information.

The presentation of the lectures as short videos also has an effect on the preparation of the lecture. Maybe owing to the fast-paced modern times, people's attention spans are getting shorter. The success of MOOCs relies — at least partly — on the very short lecture units compared to traditional university lectures of around 90 minutes. This condensed form of presentation in small chunks necessitates more precise wording from the teacher. Each phrase needs to be thoroughly planned to prevent misunderstandings which cannot be easily cleared up in contrast to classroom teaching. To handle open questions or deal with misconceptions, MOOC platforms offer space to discuss topics, ask questions, and in general interact with fellow participants and the teaching staff through a forum. This allows for deeper discussions or excursus to advanced topics, as well as to address specific aspects.

## 2 Search Engine MOOC on openHPI

We offer a Master's lecture on "Information Retrieval and Web Search" where students not only learn information retrieval concepts but also design and implement their own search engines in small teams. In addition, we offer an introduction to information retrieval and web search for high school students in a three sessions course[1]. To fill the gap between teens and students and give interested persons outside the education system the opportunity to learn about IR, we offer a MOOC on the topic on the openHPI platform.

### 2.1 openHPI

The openHPI MOOC platform[2] offers the interested public German-language and English-language online courses with a diverse computer science focus. While some courses target a broad audience and introduce fundamentals of computer science, other courses go into more detail of specialized, advanced topics.

[1] `https://hpi.de/open-campus/schuelerakademie/schuelerkolleg.html`
[2] `https://open.hpi.de/`

Course topics range from programming languages, mathematics, and IT law, to hardware-related topics, such as in-memory databases and mainframes.

The online learning platform is composed of videos, forums, quizzes and interactive programming environments. Figure 1 shows a screenshot of the website's video player, which is the starting point of all interactions on the platform. HPI actively conducts research on MOOCs and incorporates research results. For example, research suggests that videos should be segmented into short chunks (less than 6 minutes) and that the instructor's head should be shown together with the presentation slides to increase student engagement [12].

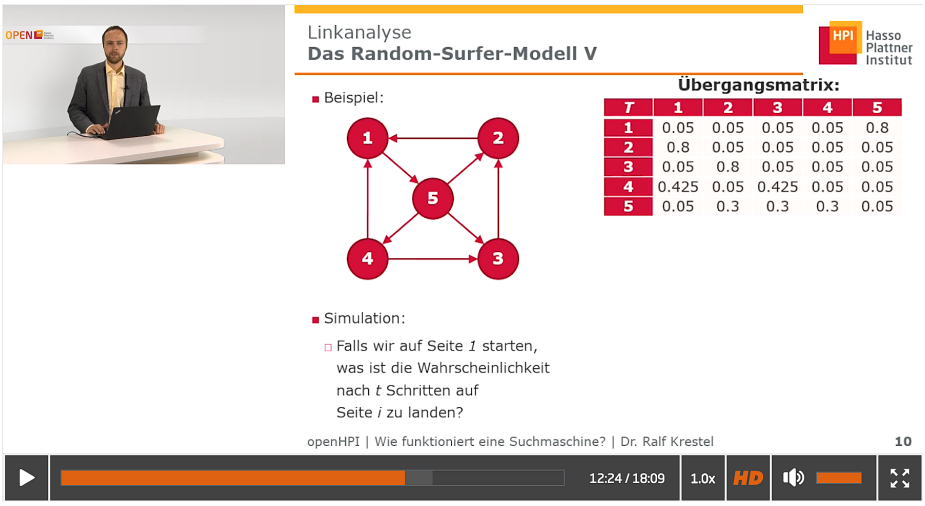


**Fig. 1.** Screenshot of the course platform website's video player.

After watching the video of a learning unit, students can ask questions or discuss the unit's content in the forum. Furthermore, they can evaluate their learning achievement with quizzes. These quizzes are composed of multiple choice questions and multiple answer questions. In the former, students need to choose the single right answer out of four possible answers, whereas in the latter, students need to choose multiple right answers out of four.

At the end of each course, there is a final exam. A record of achievement is issued to those students who have earned more than 50% of the maximum number of points for the sum of all graded assignments. A confirmation of participation is issued to those who have completed at least 50% of the course material.

### 2.2 Search Engine MOOC

The standard MOOC courses run for six weeks and cover content equivalent of a two hours per week semester course. Since we did not want to transfer our

**Table 1.** The two-week course covers a broad topic range in 17 videos.

| 1st Week | 2nd Week |
|---|---|
| 1. Introduction | 1. Search queries and user interaction |
| 2. History of information retrieval | 2. Interaction: query processing |
| 3. Text processing | 3. Interaction: query refinement |
| 4. Index construction | 4. Interaction: search engine result pages |
| 5. Ranking: Boolean retrieval | 5. Web search |
| 6. Ranking: vector space model | 6. Crawling |
| 7. Ranking: factors | 7. Social networks |
| 8. Evaluation in information retrieval | 8. Link analysis |
| | 9. New tasks and applications |

Master's lecture but to develop an introduction course without any requirements for the participants, we chose the shorter two week format. Our course "How Do Search Engines Work?" started in May 2017 with over 4000 registered participants. In 17 short videos of about twelve minutes each, we teach simple concepts of information retrieval and web search. While our videos are considerably longer (average length of 12 min) than the recommended six minutes [12], they are shorter than the videos of other openHPI courses, which was positively noted by many participants.

The content covered is inspired by university Master courses on information retrieval, but stays on a very high level without introducing algorithmic details. Students learn, for example, how a search engine is built, which process is started when searching for something, and according to what criteria the results are listed. Table 1 lists the course topics per week. Because of the student's different levels of familiarity with mathematical proofs and probability theory, the online course does not cover the details of iterative computation of PageRank, index compression, or probabilistic information retrieval.

After each video, students can take a short, optional quiz as a selftest. This quiz can be repeated as often as wanted. Figure 2 exemplifies a selftest quiz after a learning unit about link analysis and the PageRank algorithm. To answer this question, students need to rank the nodes in this graph by their PageRank value. For this, it is not necessary to compute exact values but it is sufficient to have understood the general mechanism behind PageRank.

At the end of the two-week course, there is an exam with 16 questions. The exam time is limited to one hour and once started, cannot be paused or restarted. Neither the selftest quizzes nor watching the videos are needed as qualification to take the exam. However, for preparation, students can recap the course content with randomly sampled sets of selftest quizzes. An advantage of this form of learning is the instant feedback. Immediately after submitting the quiz, students get their evaluation together with recommendations of videos they may want to revise.

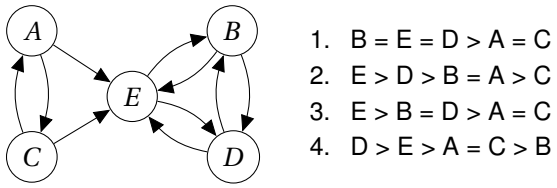


**Fig. 2.** Exemplary selftest question: Which of the node lists is ordered by PageRank with regard to the graph given on the left?

**Table 2.** In an optional survey, most students indicated to have enrolled for extended vocational training, whereas only a minority chose academic training as their reason for enrollment.

| | Vocational Training | Spare Time/Interest | Academic Training | Other |
|---|---|---|---|---|
| #Students | 700 | 636 | 72 | 54 |

## 3 Course Participants

Compared to regular information retrieval courses at university, the audience of this introductory MOOC was very heterogeneous. Furthermore, there were no pre-requisites for this course. While students indicated different reasons for their enrollment, the majority took the course for extended vocational training as shown in Table 2. In total, 4458 students enrolled, of whom 1135 watched the last video of the course and of whom 698 took the exam.

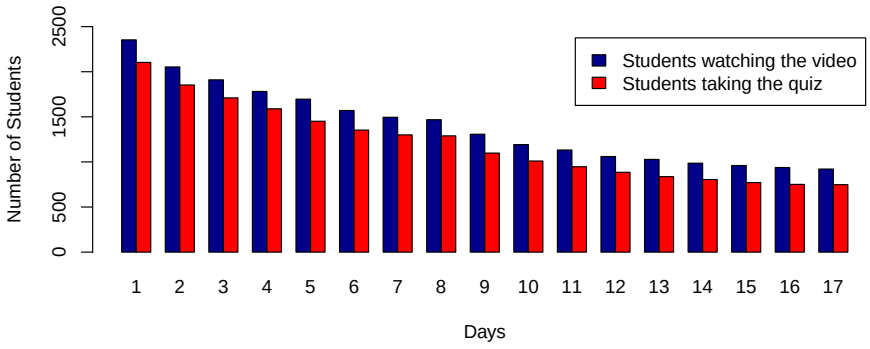


**Fig. 3.** The number of students watching the video and taking the quiz per course day falls from over 2300 to less than 1000.

### 3.1 Statistics

Of 4500 enrolled students, 2500 participated in the course (watched at least one video) and 883 took the final exam. Of these 883 students, 19% were female and 81% were male. The data basis for the following statistics are only those students, who participated in the exam. Figure 4 shows the distributions of their age and highest degree. In Figure 5, we visualize the distribution of the number of points achieved in the final exam grouped per highest degree. Each group covers a wide range of achieved points and interestingly, the median number of exam points supports the assumption that degree and number of exam points correlate. In addition, we analyze the correlation of the number of exam points and the number of visited videos and quizzes. Under the assumption of a linear correlation and according to a linear regression, we find a significant correlation of these variables. A visited video or quiz correlates with an increase by 0.2 points in the final exam.

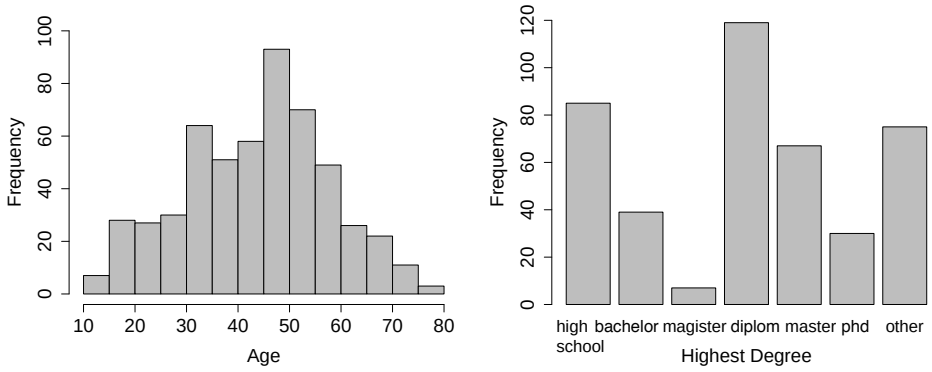


**Fig. 4.** Distributions of age and degree among 883 students who took the final exam

### 3.2 User Participation

A key advantage of MOOCs is the larger number of course participants that can discuss the learning units together. In 72 different threads, students asked questions about the course content or delved into a subject with discussions. While there were 94 answers to questions and additional 78 comments to questions and answers, only a third of these replies came from the teaching team. Two-third of the replies to questions came from the course participants themselves. Research has shown that a successful outcome is not correlated with teaching team interaction in the forum [33]. In total, there were 358 posts in the forum and 6479 views of these posts. This confirms findings by [2]: they reveal usage peaks on weekends (spare time available) and that over 90% of forum activity are passive views and only 10% are active posts. Discussed topics covered various types:

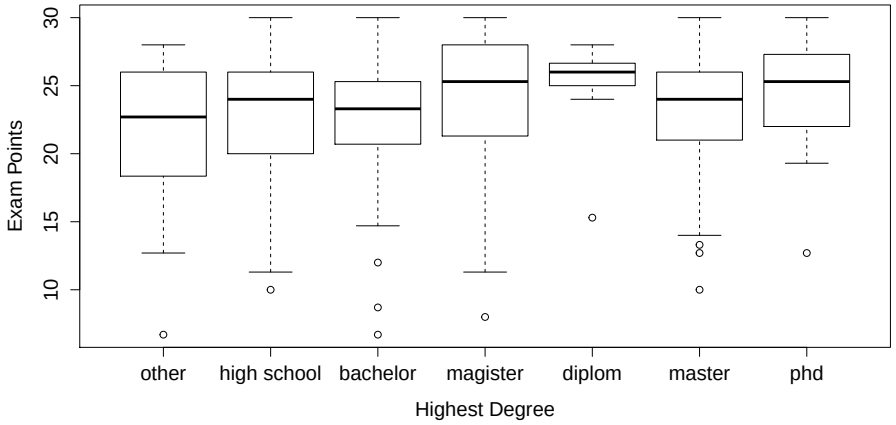


**Fig. 5.** Quartiles of achieved exam points grouped by highest degree

from technical questions, to ethical issues, and personal anecdotes. One of the most discussed topics was the bag of words model, where participants posted different examples for texts and their respective bag of words. They discussed how lemmatization and stemming can change the bag of words and consequently change also the results of search queries. But also non-technical discussions arose, e.g. about the German plural form of the word "index".

A different discussion focused on alphabetical subject catalogs and the difference to tags and keywords. An older participant asked: "Is it really true that most people don't know alphabetical subject catalogs?". While some participants were indeed not aware of library catalogs ("I know these [alphabetical subject catalogs] only from databases and electronic systems."), others remembered public libraries with index cards ("As high school student and later on as university student, catalogs were quite common for me. After watching this video, I am feeling old." or "When I was a student (70s/80s), the introductory week for new students included getting to know the library and searching for books."). Luckily, experts in the field joined the discussion: "I am librarian and of course tagging is still used. However, most people use keyword search in electronic databases and do not notice tagging."

## 4 Research Challenges on MOOC Platforms

In addition to their educational purpose for course participants, MOOC platforms provide various research opportunities. The vast amount of data gives insights into the learning process. Under the umbrella of e-learning, learning analytics, and educational data mining a research community has formed looking at various aspects of MOOCs and trying to improve the learning experience.

**Analyzing MOOCs:** An initial study of the data generated by MIT's first MOOC reports on characteristics of the students and their use of course resources [2]. Insights into a recommender systems MOOC [21] and a German database MOOC [28] confirm the findings. Also, enrollment numbers and success rates can be analyzed in detail [24].

**Infrastructure/Platform** Guo et al. study how MOOC video production affects student engagement [12]. Also other elements were the focus of research work, e.g. the forum and whether gamification is beneficial [5].

**Tools for Teachers** Designing and evaluating tools to support teachers is important. Helpful in this context is also to know more about your (potential) participants. Chen et al. match course participants with their profiles in several online social networks [4]. Thereby, they are able to compare course topics with job titles of participants. As a result, teachers learn more about students and their educational needs. Similarly, the effectiveness of certain tools needs to be evaluated, e.g. whether instructor involvement in forums has any impact on student outcomes [33]. Also prediction of student success or assessment of progress needs to be monitored, e.g. by analyzing natural language texts [6].

**Student Behavior Modeling** Understanding students' behavior is necessary to ensure a good learning experience. This starts with recommending or predicting course enrollment [27]. Once students are enrolled, keeping them motivated and helping them learn is most important. Drop-out rates are overall still high and therefore predicting drop-outs can help to intervene at the right point to keep students interested [18]. During courses, drop-out prediction can also help understanding at which point in the course users discontinue [26]. Analyzing discussion forums with respect to sentiment [34] revealed that the daily drop-out rate and sentiment expressed in forums correlate significantly. These students-at-risk should be identified and supported [14]. In addition to forums, clickstream data can be used to detect changes in student behavior [30].

**Collaborative Learning** The social component of MOOCs plays an important part for successful learning. Peer grading is very important for large courses and the quality of peer grading can be increased by motivation [22]. The authors show that students can be motivated to put more effort into peer grading if they are confronted with the effort that other graders put into it and if they are asked to evaluate other graders' efforts. Recently, the social connections between course participants and their influence on peer assessment have been studied [3]. Social comparison can be used to increase peer pressure to complete a course and reduce drop-out rates [7]. In forums, answers to questions can be ranked based on helpfulness automatically to reduce the load for other users to find the correct or best answer [17].

**Personalized Learning** The future of MOOCs is probably the personalization of the learning process. Individual participants have different learning strategies, speed, and learning preferences. This can be a huge advantage over traditional classroom teaching. First approaches for adaptive learning are already tested [32]. Further, the consideration of mood and emotions within

the learning process can help to reduce drop-out rates. Affective learning is one approach to this end [31].

## 5 Teaching Information Retrieval

Before the age of the MOOCs, e-learning environments existed that were also used for teaching information retrieval. Henrich and Morgenroth [15] report on their experience with using different e-learning scenarios in the context of IR. Similar to our experience, the forum was an important mean of communication among students and between students and faculty. We also offer the possibility for students to perform self-tests in the form of multiple choice questions after each lesson. With respect to the learning material, we provide the slides as PDF-documents along with the videos. Regarding the ordering of topics in the syllabus, we first introduced classical information retrieval concepts and in a second part elaborated on the specific Web IR challenges, as recommended by Mizzaro [25]. Henrich and Stieber [16] further analyzed IR e-learning courses along two dimensions: degree of interaction (e-learning vs. blended learning) and main media type (text-based vs. recording-based).The results indicate that all combinations can work out and other factors are more important for success, such as a clear teaching concept tailored towards the specific target audience, or active participation of students and lecturers in forums. Other papers [1, 10, 23] discuss IR curriculums and possible accompanying practical exercises for full term IR courses. Kauchak [20] reports about his experience with a course-long project consisting of the development of a search engine.While this is something that we also do for our regular IR courses, this is clearly not feasible for a short-term, introductory MOOC course. Also alternative teaching methods (inquiry-based learning) were applied to full term IR courses [19], supporting the learning-by-doing paradigm. Many related work appeared in two workshops 2007 and 2008 with the title "teaching and learning in information retrieval". In addition, in 2011 a book [9] with the same title was published covering many aspects of IR teaching in various chapters.

## 6 Lessons Learned

Some students missed course content about more advanced use of search engines and more advanced search operators. Beyond the topics of our course, students wanted to learn about search engine optimization or online reputation management for search engine results. Several students asked for practice-oriented hands-on exercises. This feedback was retrieved also for other courses on the openHPI platform and openHPI is now experimenting with concrete exercises in server-based training environments [11]. For example, the course "Data Management with SQL" incorporates practical tasks where students create SQL queries for MySQL [28].

Regarding mathematical backgrounds of the presented models, several students praised comprehensive explanations without too much detail for this short

two-week course. They aimed at a general understanding of search engines and preferred concrete examples. The exemplary indexing or ranking of documents and the expansion of queries was perceived more informative than theoretical evaluation measures, such as MAP or NDCG. Although the two course weeks have ended, the course material stays available. For this reason, we expect the enrollment number to increase further.

There seems a demand for short introductory courses in IT especially from older participants. To allow flexible learning, a single video shouldn't exceed 10 minutes in length and the video length variance should be small for individual planning reasons. Also longer MOOC courses (six weeks) on IR with more details are possible to increase the level of detail and add some more practical examples and homework tasks.

## References


1. Blank, D., Fuhr, N., Henrich, A., Mandl, T., Rölleke, T., Schütze, H., Stein, B.: Information retrieval: Concepts and practical considerations for teaching a rising topic. Datenbank-Spektrum 9(29), 30–41 (2009)
2. Breslow, L., Pritchard, D.E., DeBoer, J., Stump, G.S., Ho, A.D., Seaton, D.T.: Studying learning in the worldwide classroom: Research into edx's first mooc. Research & Practice in Assessment 8 (2013)
3. Chan, H.P., King, I.: Leveraging social connections to improve peer assessment in moocs. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 341–349. International World Wide Web Conferences Steering Committee (2017)
4. Chen, G., Davis, D., Lin, J., Hauff, C., Houben, G.J.: Beyond the mooc platform: gaining insights about learners from the social web. In: Proceedings of the 8th ACM Conference on Web Science. pp. 15–24. ACM (2016)
5. Coetzee, D., Fox, A., Hearst, M.A., Hartmann, B.: Should your mooc forum use a reputation system? In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing. pp. 1176–1187. CSCW '14, ACM, New York, NY, USA (2014)
6. Crossley, S., Liu, R., McNamara, D.: Predicting math performance using natural language processing tools. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference. pp. 339–347. LAK '17, ACM, New York, NY, USA (2017)
7. Davis, D., Jivet, I., Kizilcec, R.F., Chen, G., Hauff, C., Houben, G.J.: Follow the successful crowd: Raising mooc completion rates through social comparison at scale. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference. pp. 454–463. LAK '17, ACM, New York, NY, USA (2017)
8. Dillahunt, T., Wang, B., Teasley, S.: Democratizing higher education: Exploring mooc use among those who cannot afford a formal education. The International Review of Research in Open and Distributed Learning 15(5) (2014)
9. Efthimiadis, E., Fernández-Luna, J.M., Huete, J.F., MacFarlane, A.: Teaching and learning in information retrieval, vol. 31. Springer Science & Business Media (2011)
10. Fernández-Luna, J.M., Huete, J.F., MacFarlane, A., Efthimiadis, E.N.: Teaching and learning in information retrieval. Information Retrieval 12(2), 201–226 (Apr 2009), https://doi.org/10.1007/s10791-009-9089-9

11. Grünewald, F., Mazandarani, E., Meinel, C., Teusner, R., Totschnig, M., Willems, C.: openhpi: Soziales und praktisches lernen im kontext eines mooc. In: DeLFI. pp. 143–154 (2013)
12. Guo, P.J., Kim, J., Rubin, R.: How video production affects student engagement: An empirical study of mooc videos. In: Proceedings of the First ACM Conference on Learning @ Scale Conference. pp. 41–50. L@S '14, ACM, New York, NY, USA (2014)
13. Hansen, J.D., Reich, J.: Democratizing education? examining access and usage patterns in massive open online courses. Science 350(6265), 1245–1248 (2015)
14. He, J., Bailey, J., Rubinstein, B.I.P., Zhang, R.: Identifying At-Risk Students in Massive Open Online Courses. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA. pp. 1749–1755 (2015)
15. Henrich, A., Morgenroth, K.: Information retrieval as elearning course in germanlessons learned after 5 years of experience. In: Proceedings of the First International Workshop on Teaching and Learning in Information Retrieval, London, UK (2007)
16. Henrich, A., Sieber, S.: Blended learning and pure e-learning concepts for information retrieval: experiences and future directions. Information Retrieval 12(2), 117–147 (Apr 2009), https://doi.org/10.1007/s10791-008-9079-3
17. Jenders, M., Krestel, R., Naumann, F.: Which answer is best?: Predicting accepted answers in mooc forums. In: Proceedings of the 25th International Conference Companion on World Wide Web. pp. 679–684. International World Wide Web Conferences Steering Committee (2016)
18. Jiang, S., Williams, A.E., Schenke, K., Warschauer, M., O'Dowd, D.K.: Predicting MOOC performance with week 1 behavior. In: Proceedings of the 7th International Conference on Educational Data Mining, EDM 2014, London, UK, July 4-7, 2014. pp. 273–275 (2014)
19. Jones, G.J.: An inquiry-based learning approach to teaching information retrieval. Information Retrieval 12(2), 148–161 (2009)
20. Kauchak, D.: A course-long information retrieval project. In: First AAAI Symposium on Educational Advances in Artificial Intelligence (2010)
21. Konstan, J.A., Walker, J.D., Brooks, D.C., Brown, K., Ekstrand, M.D.: Teaching recommender systems at large scale: Evaluation and lessons learned from a hybrid mooc. ACM Trans. Comput.-Hum. Interact. 22(2), 10:1–10:23 (Apr 2015)
22. Lu, Y., Warren, J., Jermaine, C.M., Chaudhuri, S., Rixner, S.: Grading the graders: Motivating peer graders in a MOOC. In: Proceedings of the 24th International Conference on World Wide Web. pp. 680–690. ACM (2015)
23. McCown, F.: Teaching web information retrieval to undergraduates. In: Proceedings of the 41st ACM Technical Symposium on Computer Science Education. pp. 87–91. SIGCSE '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1734263.1734294
24. Meinel, C., Willems, C., Renz, J., Staubitz, T.: Reflections on enrollment numbers and success rates at the openhpi mooc platform. Proceedings of the European MOOC Stakeholder Summit pp. 101–106 (2014)
25. Mizzaro, S.: Teaching of web information retrieval: Web first or ir first. In: Proceedings of the first international workshop on teaching and learning in information retrieval (TLIR 2007). pp. 50–54 (2007)
26. Nagrecha, S., Dillon, J.Z., Chawla, N.V.: Mooc dropout prediction: Lessons learned from making pipelines interpretable. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 351–359. International World Wide Web Conferences Steering Committee (2017)

27. Nam, S., Lonn, S., Brown, T., Davis, C., Koch, D.: Customized course advising: investigating engineering student success with incoming profiles and patterns of concurrent course enrollment. In: Learning Analytics and Knowledge Conference 2014, LAK '14, Indianapolis, IN, USA, March 24-28, 2014. pp. 16–25. ACM (2014)
28. Naumann, F., Jenders, M., Papenbrock, T.: Ein datenbankkurs mit 6000 teilnehmern. Informatik-Spektrum 37(4), 333–340 (2014)
29. Pappano, L.: The year of the mooc. The New York Times 2(12), 2012 (2012)
30. Park, J., Denaro, K., Rodriguez, F., Smyth, P., Warschauer, M.: Detecting changes in student behavior from clickstream data. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference. pp. 21–30. LAK '17, ACM, New York, NY, USA (2017)
31. Ruiz, S., Charleer, S., Urretavizcaya, M., Klerkx, J., Fernández-Castro, I., Duval, E.: Supporting learning by considering emotions: Tracking and visualization a case study. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. pp. 254–263. LAK '16, ACM, New York, NY, USA (2016)
32. Sonwalkar, N.: The first adaptive mooc: A case study on pedagogy framework and scalable cloud architecture — part i. MOOCs Forum 1(P), 22–29 (2013)
33. Tomkin, J.H., Charlevoix, D.: Do professors matter?: Using an a/b test to evaluate the impact of instructor involvement on mooc student outcomes. In: Proceedings of the First ACM Conference on Learning @ Scale Conference. pp. 71–78. L@S '14, ACM, New York, NY, USA (2014)
34. Wen, M., Yang, D., Rose, C.: Sentiment analysis in mooc discussion forums: What does it tell us? In: Educational Data Mining 2014 (2014)
35. Zemsky, R.: With a mooc mooc here and a mooc mooc there, here a mooc, there a mooc, everywhere a mooc mooc. The Journal of General Education 63(4), 237–243 (2014)