Predicting dropouts on the successive offering of a MOOC

Massimo Vitiello[1], Simon Walk[1], Denis Helic[1], Vanessa Chang[2], and Christian Guetl[1 2]

[1] Graz University of Technology massimo.vitiello@student.tugraz.at, simon.walk@tugraz.at, dhelic@tugraz.at, cguetl@iicm.edu
[2] Curtin University of Technology vanessa.chang@curtin.edu.au

Abstract. In recent years, we have experienced the rise of e-learning and the growth of available Massive Online Open Courses (MOOCs). Consequently, an increasing number of universities has dedicated resources to the development and publishing of MOOCs on portals. A common practice for operators of such MOOCs is to re-offer the same course over the years with similar modalities and minor improvements. Such *re-runs* are still affected, as most of the MOOCs, by a low percentage of enrolled users who manage to successfully complete the courses. Hence, analyzing similar MOOCs can provide valuable insights to better understand the reasons of users for dropping out and potentially can help MOOCs' administrators to better shape the structure of the courses to keep users engaged. To that end, we analyze two successive offerings of the same MOOC, created by Curtin University and published on the edX platform. We extract features for our prediction experiment to detect dropouts, considering two different metrics: (i) the percentage of users active time and (ii) the initial week after users first interaction with the MOOC. We train a Boosted Decision Tree classifier with the extracted features from the original run of the MOOC and predict dropouts on its re-run. Furthermore, we identify a set of features that likely indicates whether users will drop out in the future or not. Our results indicate that users interacting with particular tools at the very beginning of a MOOC are more likely to successfully complete the course.

1 Dropouts in MOOCs

Online learning enormously changed over the past years. The Internet became the channel on which a variety of new types of learning methodologies materialized. Massive Online Open Courses (MOOCs) emerged as the natural solution to offer distance education. The advantages of MOOCs are various. They are *Massive*, in the sense that a potentially unlimited audience can enroll; *Online*, as all that is needed, is an internet connection, without any geographical limitation; *Open*, since the majority of them have no enrollment costs, nor do they require particular prerequisites; a *Course*, as their structures resemble traditional lectures, with assignments and exams. [9, 10] Despite these advantages, the expectations MOOCs carried with them have not yet been completely reached.

Nearly all MOOCs suffer from high dropout rates. While the number of users who enroll is high, the percentage of those that successfully complete the course is very low (generally lower than 10%) [7]. High dropout rates are a problem not only for single runs of MOOCs but are also present in successive offerings of the same course. We refer to the same offering of a MOOC, happening at a later point in time, as *re-run*. These are characterized by having structures, topics, and schedules similar to those of the original course. Therefore, lower efforts are required when organizing successive re-runs. Furthermore, content creators can make use of previous users' feedback to modify and reshape the re-run, in order to better meet users' goals and expectations, while generally enhancing the overall learning experience. The investigation and comparison of re-runs of the same MOOC can help us to increase our understanding of the learning style of the users and how to support them, with the clear goal of mitigating dropout rates. In this paper, we experiment with early dropout detection on MOOC re-runs. Particularly, we analyze 2 MOOCs offered on edX by Curtin University (Perth, Western Australia)[3], the second of which is the re-run of the first one. First, we investigate a varying percentage of users total active time. Second, we focus on the first week of interactions of each user. This allows us to verify if users' behavior during the initial stage of the course is a strong indicator of their outcome and if this is also true for re-runs. Furthermore, we identify which features are the most valuable to correctly predict if users will drop out or not.

2 Related Work

Gütl et al. [5] analyzed survey answers of users who dropped out, across MOOCs offered by Universidad Galileo. They proposed an Attrition Model for Open Learning Environment Setting (AMOES), which builds up and extends the Funnel of Participation Model [2], and split the attrition into healthy and unhealthy. Within healthy attrition, they identify 3 subgroups according to users' goals, expectations, and reasons to drop out. Particularly, users are classified as either *Exploring User*, *Content Learner* or *Restricted Learner*.

Coffrin and colleagues [3] studied two MOOCs developed at the University of Melbourne. Principles of Macroeconomics was an introductory course with the material available in a linear structure over its 8 weeks duration. Discrete Optimization was a self-paced graduate level course, which lasted 9 weeks. The authors used a linear regression model and analyzed the relation between users' final grade and their interactions during the first two weeks of the MOOCs. They also used State Transition Diagrams to discover similarities across the users of the MOOCs, by visualization of assignment and weekly video interactions.

Kloft and al. [8] experimented with weekly dropouts classification using Support Vector Machines (SVM) on a 12 weeks MOOC with an 81.4% dropout rate offered on Coursera. The authors used cumulative features (number of interaction, number of view of each page of the course) and technical ones (browser, OS, number of screen pixel). They compared a trivial baseline (always predicts one or the other class) to the performance of SVM, which showed higher accuracy.

[3] https://www.edx.org/school/curtinx

Amnueypornsakul and colleagues [1] experimented with dropout classification using SVM on a MOOC with roughly 30,000 enrolled students. The authors used quiz related and activity related features, verifying by ablation analysis that the two sets are both important for the prediction task. Furthermore, they noticed that the class imbalance and the presence of users with few interactions (Inactive) complicate the classification task.

In our previous work [13] we experimented with dropout prediction across 5 MOOCs offered by Universidad Galileo on the university portal Telescopio. They derived features from users' sessions and also considered the number of time a tool was used as a measure. They analyzed the MOOCs using SVM and K-Means as classifiers and tested different combinations of features. In their results, K-Means always fell behind SVM and the prediction was improved by different combinations of features. In 2017 We refined our approach in [12] and developed a general classifier for dropout detection across different MOOC platforms.

Teusner et al. [11] analyzed 3 iterations of the MOOC "In-Memory Data Management (IMDM)" offered on the onpenHPli platform, hosted and developed by the Hasso Plattner Institute of Potsdam. The MOOC was held in English and was intended for learners with a business background and academics. While the content of the first two interactions barely differed, the third offering was improved according to user feedback. Some units were reshaped to ease understanding, and about 60% of all videos were modified. The authors concluded that future interactions with stable material attracted a wider audience with low effort, using the forum for content-based communications can help to promote users' engagement and content-related feedback should be introduced and addressed as fast as possible.

3 Materials and Methods

3.1 Datasets

Our dataset consists of the original offering of a MOOC, referred to as *MOOCC1*, and the first of its re-runs, coded as *Re-Run1*. Both MOOCs were created by Curtin University and were available on edX[4]. The original offering *MOOCC1*

[4] https://www.edx.org/school/curtinx

Table 1. Summary of the MOOC and the re-run. *Active* users are the part of *Enrollments* that conducted at least more than one interaction. The second class of users is the *Inactive* one. In the column *Dropouts* we indicate the number of users that drop out in relation to the *Active* users and in relation to the *Enrollments* (in brackets). Analogously, the dropout rates are relative to *Active* users and to the *Enrollments*.

MOOC	Enrollments	Active	Inactive	Completers	Dropouts	Dropout Rate
MOOCC1	21948	13396	8552	1500	11896 (20448)	89% (93%)
Re-Run1	10368	5932	4436	208	5724 (10160)	96% (98%)

was available online during the second semester of 2015, while the re-run *Re-Run1* was available online between April and May 2016. Both offers had no entry prerequisites. However, users needed to have access to YouTube in order to watch the independently created video content. Regular activities, such as polls, questions, and discussion board tasks were integrated into the course content. The course syllabus consisted of a total of four modules, each estimated to require a time commitment of two hours per week. An extra introductory module and a course wrap up module completed the course calendar. To complete each of the four main modules, participants needed to complete an activity and a quiz, with the quizzes being an extension of activities. Therefore, engaging in the activities helped participants to answer the questions in the quizzes. Each quiz accounted for 25% of the final grade, with a Certificate of Achievement issued to participants with an overall score of equal or greater than 70%. The team of instructors consisted of an associate professor and a teaching assistant. The two offers were for the larger part similar to each other regarding contents and activities, with *Re-Run1* undergoing only some minor changes. Table 1 reports a summary of the enrollments and completions of the original MOOCs and of its re-run. As the column *Enrollments* reports, *MOOCC1* has a total of 21,948 enrolled users and its re-run, *Re-Run1*, counts 10,368 enrolled users. Within the enrolled users, we distinguish users that enroll and leave the MOOCs without engaging any further (i.e., *Inactive* users), from those who have more than the simple enrollment interaction (i.e., *Active* users). *Completers* are users that successfully completed the MOOCs, while *Dropouts* are those who failed to do so. Overall, the dropout rates are never lower than 89%.

Both offers are structured in a self-paced manner and are organized in two phases. During the first phase, users can only access the course main page and enroll. At this stage, the course's material is not available, and only a limited number of interactions is possible. This initial phase lasts for roughly 2 months for both MOOCs. At the beginning of the second phase, the course material is uploaded all at once and users can engage at their own pace. Enrollment is possible during the second phase as well. After the official end of the MOOCs, users still can register and interact with the MOOC, but, in this case, they can not obtain a certificate as the course is already officially over. Due to these settings, we consider only enrollments happening before the official end of the MOOC. Furthermore, we also discard interactions happening before the course's official start. Such interactions do not represent users learning style because the course material is not available yet. Both MOOCs also included a course forum where users can post and discuss.

3.2 Experimental Setup

We extract a set of features to describe each user of the MOOCs from Curtin University. First, we calculate a set of time-based features that build upon the concept of sessions. A session is a set of chronologically ordered interactions, in which each interaction happens within a certain timespan from the previous and the next one. Particularly, we use a threshold of 30 minutes and calculate sessions for all users. Following this concept of sessions, we define the following features:

Sessions as the total number of users' sessions; *Requests* as the total number of interactions per user; *Active Time* is the total time users interacted with the MOOC (as sum of all sessions' duration); *Days* indicates the number of days during which users interacted at least once with the MOOC. Furthermore, we compute 4 averaged features; *Timespan Clicks* is the average timespan between two consecutive clicks in the same session (averaged over all sessions); *Session Length* as *Active Time* divided by *Sessions*; *Session Requests* as *Requests* over *Sessions*; *Day Requests* is *Requests* divided by *Days*. Moreover, we exploit the detailed edX logs, to identify the type of event triggered and the particular tool each interaction refered to. Curtin University's MOOCs includes 6 specific tools[5]; *Course Navigation*, *Video*, *Problem*, *Poll & Survey*, *Bookmark* and *Discussion Forum*. For each of these tools, a set of events indicates the particular action that took place. The list of events for each tool, together with the session related features we calculated, are reported in Table 2. The *Video* tool, is the only one that allows distinction between Browser and Mobile (through the edX mobile application) triggered interactions. To categorize these two sources, we create an additional tool, filtering out Mobile interactions from *Video*, and we name this new tool *Video Mobile*. We create a feature for each event, counting the number of times each of these was triggered by users' interactions. Secondly, we define two different approaches and calculate the features according to these. First, we consider a varying percentage of users total active time. Particularly, we consider

[5] The complete list of edX's events is available at http://edx.readthedocs.io

Table 2. Summary of Tools and their events. In the first column we list the name of the tools and in the second column we report the list of events referring to that particular tool.

Tools	Events
Session Related	Sessions, Requests, Active Time, Days, Timespan Clicks, Session Length, Session Requests, Day Requests
Main Page Links	About, Faqs, Home, Instructor, Progress, StudyAtCurtin
Course Navigation	TabSelected, PreviousTabSelected, NextTabSelected, LinkClicked, OutlineSelected
Video	CaptionHidden, CaptionShown, LanguageMenuHidden, LanguageMenuShown, Loaded, Paused, Played, PositionChanged, SpeedChanged, Stopped, TranscriptHidden, TranscriptShown
Video Mobile	CaptionHiddenM, CaptionShownM, LanguageMenuHiddenM, LanguageMenuShownM, LoadedM, PausedM, PlayedM, PositionChangedM, SpeedChangedM, StoppedM, TranscriptHiddenM, TranscriptShownM
Problem	Check, CheckFail, FeedbackHintDisplayed, Graded, HintDisplayed, Rescore, RescoreFail, Reset, ResetFail, Save, SaveFail, SaveSuccess, Show, ShowAnswer
Poll & Survey	PollSubmitted, PollViewResults, SurveySubmitted, SurveyViewResults
Bookmark	Accessed, Added, Listed, Removed
Discussion Forum	CommentCreated, ResponseCreated, ResponseVoted, Searched, ThreadCreated, ThreadVoted

all interactions within the first 1% to 100% of the total active time (per user) and call this setting *Scaled Time*.

As a second approach, we calculate features considering the first 7 days after a users' first interaction. We name this setting *Days*. To overcome the class imbalance problem [6], we adjust the class distribution of our MOOCs by randomly oversampling the smaller class. Randomly picked examples from the smaller class are added until Completers and Dropouts have the same number of samples in our dataset. Each of the different approaches builds the foundation for a classification experiment that we run using Boosted Decision Trees [4]. This ensemble classifier combines a set of single decision tree into a single classifier. For each model, the misclassified examples get a higher weight, so the next decision tree focuses more on correctly predicting these.

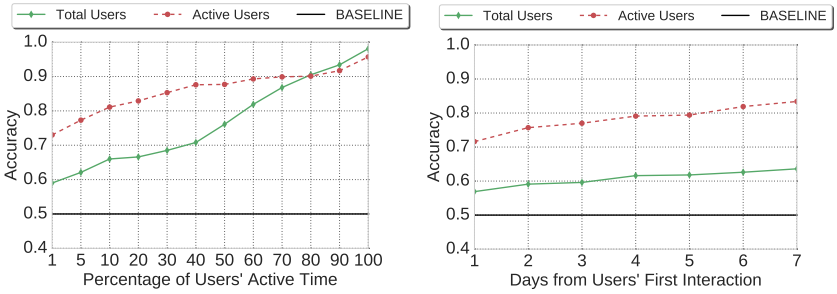
We run classification experiments with Boosted Decision Trees for both approaches using two different set of users as input. First, we consider the all *Enrollments* and then only the *Active* users, as indicated in Table 1. We evaluate our experiments using accuracy, which is calculated as the number of correctly predicted examples divided by the total number of predicted examples. Therefore, this measure can assume values between 0 and 1; a value of 0 indicates that all examples have been wrongly classified, while a value of 1 means that all examples have been correctly predicted.

As a final analysis, we investigate the importance of our features in the classification task. Boosted Decision Trees also provide a weight for each used feature. Their weight represents the number of times a feature is used to split the data across every single decision tree. Thus, the higher the weight, the more precise the obtained split. We explore the ranking of the features for both metrics when the input consists only of *Active* users.

4 Results & Discussion

Figures 1(a) and 1(b) report the results of the classification for the two approaches. The x-axes indicate the days after a user's first interaction for the Days experiment and the percentage of a users' active time for the Scaled Time experiment. For both figures, the y-axes indicate accuracy and are bounded between 0.4 and 1. We also plot the baseline as a solid black horizontal line at 0.5. The baseline represents the performances of a classifier that randomly predicts a class. It is a lower bound; classifiers with an accuracy under the baseline are no better than random prediction. Green lines refer to experiments using all enrolled users (Enrollments), while the ones in red consider only the active users (Active).

For the Days experiment reported in Figure 1(b), the accuracy, when only the active users are considered, is always increasing the more days are considered. It is never lower than 0.7 and increases over 0.8 when all 7 days after users' first interaction are considered. This indicates how the first week of user interactions already represent a good indication about which users will eventually drop out. The accuracy, when all enrolled users are considered, floats around 0.6. This indicates that in this case, the feature set does not characterize the two classes.



(a) Scaled Time (b) Days

Fig. 1. Dropout prediction results on Re-Run1. Each subfigure depicts the accuracy results for the two proposed approaches. Figure 1(a) reports the results for the Scaled Time approach and Figure 1(b) depicts the results for the Days approach, with the x-axes indicating the considered percentage of user active time or the considered number of days after user first interaction with the MOOC respectively. The y-axes (bounded between 0.4 and 1) of each figure indicate the accuracy, with the baseline being plotted as a solid black line at 0.5. The green lines refer to the experiments in which we consider all users (cf. Enrollments in Table 1), while the ones in red represent experiments in which we analyzed only the active users (cf. Active in Table 1). Considering only active users always yields the highest accuracy, except when the considered percentage of a user's active time is larger than 80%. In this case, the green and red lines switch places, however, the overall performance of the prediction experiments only minimally differs.

It is likely that a lot of users register for the MOOCs at an early stage, or at least more than a week before the material becomes available. At this time, there are few possible interactions, and it is likely that users only come back once the material is made available. Thus, this approach has a lower overall accuracy.

The results of the Scaled Time experiment are plotted in Figure 1(a). Also, with this setting, considering only the active users is the approach that produces the highest accuracy. Again, the accuracy is never lower than 0.7 and gets as high as 0.96 when the whole users' active time is considered. When all enrolled users are taken into account, the accuracy ranges from 0.59 to 0.98, constantly increasing as the percentages of users' active time get higher. Both settings have a similar profile when these percentages get higher than 80%.

Table 3 lists the best performing features for the two approaches when the active users are considered. The first column lists the Tool and the second the specific features that belong to it. The remaining columns report the weights for the features, first for the Days and then for the Scaled Time experiments. For reasons of space, we report only some values for both approaches. Particularly, we show day 1, 4 and 7 for the First 7 Days approach and 5%, 50% and 100% of users' active time. The weights highlighted bold are the highest for that particular experiment. We see that *Progress* is always one of the features with the highest weight for both experiments. Interactions of type *Progress* refer to users accessing a dedicated page to track their scores for single problems and the current overall

course grade. Particularly, this page includes a grading chart, which reports the obtained scores on each graded assignment in the form of a bar chart. Moreover, the page also offers a panoramic overview of the whole set of problems, organized per-section and listed in the order they occur in the MOOC. The weight of this feature increases with the days and time percentage. If we extract the number of interactions of this type for both classes, we obtain a total of $17,240$ for the Completers and of $30,916$ for the Dropouts for *Re-Run1*. For *MOOCC1* we have $121,228$ interactions for the Completers and $54,768$ for the Dropouts. The class average yields 82.88 and 80.98 *Progress* interactions for the Completers, and 6.88 and 5.40 for the Dropouts of *Re-Run1* and *MOOCC1* respectively. Besides, if for the Dropouts we include also the users with only the enrollment action, the averages gets as low as 3.04 for *Re-Run1* and 2.68 for *MOOCC1*.

Hence, we find that together with having a high number of correctly solved problems, constantly monitoring the personal progress strongly indicates whether a user will drop out or not. Similarly, *ProblemCheck* becomes more significant, the more days and higher percentages of interactions are analyzed. This in-

Table 3. Feature Scores of the Days and Scaled Time Experiments. The features with the highest scores for 1, 4, and 7 days after users' first interaction with the MOOC and 5%, 50% and 100% of the active time per user are boldfaced. *Progress* is always among the features with the highest scores. *Timespan Clicks* is among the features with the highest scores for the Days experiment, while *Session Length* is one of the features with the highest scores for the Scaled Time experiment. *ProblemCheck* scores increase the more days after users' first interaction and active time per user we consider.

		Days			Scaled Time		
Tool	Feature	1	4	7	5%	50%	100%
Session Related	Timespan Clicks	**62.3**	**53.9**	**57.4**	**63.1**	28.6	10.7
	Active Time	**44.6**	**39.3**	32.9	40.6	25.2	16.4
	Session Length	36.3	38.2	28.2	**59.5**	**39.3**	**56.1**
	Requests Active Day	24.6	18.1	17.0	21.8	29.7	28.8
	Session Requests	34.0	36.0	35.8	23.5	36.0	22.7
Problem	ProblemCheck	21.4	32.2	**41.0**	28.6	**47.5**	**59.8**
	ProblemGraded	3.1	7.6	4.0	1.1	13.3	42.7
	ProblemShow	18.0	16.8	20.2	18.4	11.6	35.8
Main Page Links	Home	28.0	28.8	22.6	28.9	16.5	12.5
	Progress	**50.6**	**54.2**	**54.7**	**55.1**	**84.6**	**93.8**
	StudyAtCurtin	22.0	18.8	11.0	9.9	2.6	7.7
Course Navigation	NextSelected	22.1	13.2	11.2	21.8	29.7	14.6
	TabSelected	26.9	18.3	17.4	29.2	8.8	22.2
Video	VideoLoaded	24.6	20.3	17.7	22.4	11.4	9.0
	VideoPlayed	16.6	15.2	9.6	22.4	14.2	11.1

teraction indicates a problem being correctly checked by the system after users submitted an answer. The high scores of this feature come as no surprise, as users are likely to solve problems only after they study and learn from the course's material, that is, at a later stage during the MOOC.

Certain tools and their features never obtain significant weights, such as *Video Mobile* or *Forum Discussion*. The low weights for *Video Mobile* indicate that users mostly interact with the MOOCs using a desktop machine rather than the edX mobile application. *Poll & Survey* and *Bookmark* are rarely used tools, either due to being poorly advertised or to users not regarding them as particularly useful to complete MOOCs. It also appears that interactions within the *Forum Discussion* barely relate to Completers or Dropouts. First, it is possible that the course's structure does not require users to engage with the forum. This could be due to unchallenging courses or, more likely, due to the self-paced setting of the MOOCs. Users engage at their own pace and confront the same challenges at different times. As a consequence, the role of the forum as a real-time communication channel and as the first source of help might be limited.

5 Conclusion & Future Work

In this work, we experimented with dropout detection on a MOOC re-run offered on the edX portal by Curtin University. We train a Boosted Decision Tree classifier on the initial offering of a MOOC and predict users that will drop out on its re-run. Our results indicate that the first week of users' interactions already provide information about whether users will complete the re-run or eventually drop out. Furthermore, we evaluate the importance of each of our proposed features for the classification. We discover that the frequency users check their progress and correctly solve problems within a short period after first interacting with a MOOC, are related to users' probability of completing the MOOC. We also note that certain tools are barely used by users and, therefore, do not carry any valuable information for the prediction task. Moreover, we find that the benefits of social tools, as in our case the discussion forums, appear to be related to the way MOOCs are organized (i.e., limited benefits for self-paced MOOCs) and on the efforts they require for users to engage with them.

Analyses at tool level can represent a valuable next step. Abstracting from the particular event that took place, might help to differentiate more precisely most and less used tools, and to confirm our current findings. Using different approaches that focus on the initial users' interactions, will help learn more about users' behaviors. Analogously to interaction and click-pattern mining approaches from other domains [14–17], we plan on identifying interaction types of users by clustering users of MOOCs according to their click- and interaction patterns to improve dropout detection.

6 Acknowledgments

This work is in part supported by the Graz University of Technology, Curtin University, and the MOOC Maker Project (http://www.moocmaker.org/, Reference: 561533-EPP-1-2015-1-ES-EPPKA2-CBHE-JP). The authors particularly thank Curtin University for providing the analyzed datasets.

References

1. Amnueypornsakul, B., Bhat, S., Chinprutthiwong, P.: Predicting attrition along the way: the uiuc model (2014)
2. Clow, D.: Moocs and the funnel of participation. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge. pp. 185–189. ACM (2013)
3. Coffrin, C., Corrin, L., de Barba, P., Kennedy, G.: Visualizing patterns of student engagement and performance in moocs. In: Proceedings of 4^{th} international conference on learning analytics and knowledge. pp. 83–92. ACM (2014)
4. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics pp. 1189–1232 (2001)
5. Guetl, C., Chang, V., Hernández Rizzardini, R., Morales, M.: Must we be concerned with the massive drop-outs in mooc? an attrition analysis of open courses. In: Proceedings of the International Conference Interactive Collaborative Learning, ICL2014 (2014)
6. Guo, X., Yin, Y., Dong, C., Yang, G., Zhou, G.: On the class imbalance problem. In: Natural Computation, 2008. ICNC'08. Fourth International Conference on. vol. 4, pp. 192–201. IEEE (2008)
7. Jordan, K.: Initial trends in enrolment and completion of massive open online courses. The International Review of Research in Open and Distributed Learning 15(1) (2014)
8. Kloft, M., Stiehler, F., Zheng, Z., Pinkwart, N.: Predicting mooc dropout over weeks using machine learning methods (2014)
9. McAuley, A., Stewart, B., Siemens, G., Cormier, D.: The mooc model for digital practice (2010)
10. Rodriguez, C.O.: Moocs and the ai-stanford like courses: Two successful and distinct course formats for massive open online courses. European Journal of Open, Distance and E-Learning 15(2) (2012)
11. Teusner, R., Richly, K., Staubitz, T., Renz, J.: Enhancing content between iterations of a mooc–effects on key metrics (2015)
12. Vitiello, M., Walk, S., Chang, V., Hernández, R., Helic, D., Guetl, C.: MOOC dropouts: A multi-system classifier. In: 12th European Conference on Technology Enhanced Learning, EC-TEL 2017, Tallinn, Estonia. pp. 300–314 (2017)
13. Vitiello, M., Walk, S., Hernández, R., Helic, D., Gütl, C.: Classifying students to improve mooc dropout rates pp. 501–508 (2016)
14. Walk, S., Espín-Noboa, L., Helic, D., Strohmaier, M., Musen, M.A.: How Users Explore Ontologies on the Web: A Study of NCBO's BioPortal Usage Logs pp. 775–784 (2017)
15. Walk, S., Singer, P., Espín-Noboa, L., Tudorache, T., Musen, M.A., Strohmaier, M.: Understanding how users edit ontologies: Comparing hypotheses about four real-world projects. In: 14th International Semantic Web Conference, USA, October, 2015. pp. 551–568 (2015)
16. Walk, S., Singer, P., Strohmaier, M.: Sequential action patterns in collaborative ontology-engineering projects: A case-study in the biomedical domain. In: 23rd ACM International Conference on Information and Knowledge Management, CIKM, Shanghai, China, 2014. pp. 1349–1358 (2014)
17. Walk, S., Singer, P., Strohmaier, M., Tudorache, T., Musen, M.A., Noy, N.F.: Discovering beaten paths in collaborative ontology-engineering projects using markov chains. Journal of Biomedical Informatics 51, 254–271 (2014)