4th International Conference on Higher Education Advances (HEAd'18)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/HEAd18.2018.7987

Increasing Peer Review Quality in Online Learning Systems

Luckner, Naemi; Purgathofer, Peter; Fitzpatrick, Geraldine
Human Computer Interaction Group, Institute of Visual Computing and Human-Centered Technology, Wien.

Abstract

Lecturers face an on-going struggle to keep up-to-date with their students' learning progress in large university courses. This hurts especially when it comes to identifying and supporting the diverse needs of each individual student. One way to approach this challenge is to introduce peer reviewing as a means to provide students with individual feedback throughout the semester. However, the quality of feedback written by peers can vary immensely and some students intentionally avoid putting work into writing reviews. We addressed these issues by calculating a Review Karma (RK), a value indicating how helpful students are in giving feedback to their colleagues and in helping them improve. While this approach shows much promise, especially in identifying different groups of students and enhancing their learning experience, we also identified trends that negatively impact the way students approach reviewing and provide their honest opinions of their colleagues work. The main contributions of this paper are the design of and lessons learned from the introduction of the RK and its initial evaluation via a survey.

Keywords: *Peer Review; Review Quality; Peer Feedback; Peer Evaluation*

1. Introduction

Peer review is a widely established process to give and receive personalised feedback. For peer review to work well, reviewing has to be taken seriously and the feedback needs to be constructive and helpful. This, however, proves to be challenging to get across to students of large university courses, many of whom see reviewing as a chore to quickly get out of the way for 'more important' work. We are faced with the challenge to improve the overall review quality, combat negligent review practices and get students to value reviewing as an important activity in their course work.

Since 2013, we have been exploring the use of double blind peer reviewing in large university courses (>500 participants). Using an in-house system [blinded ref], we facilitated approximately 45.000 peer reviews per semester in two courses between 2013 and 2017. In these courses, an essential part of students' course work consists of writing peer reviews. These reviews are highly structured, usually consisting of a set of questions that are specific to the elaboration and cannot be answered with simple one-word replies.

To combat negligent reviewing practices, we introduced a number of measures over the years including better justifying reviewing to the students by explaining the value for everybody, giving students the opportunity to send anonymous feedback to the review author, and increasing the weight of individual review quality in the final grade. This paper discusses yet another measure we introduced in order to bring students to strive for a higher standard in writing their reviews, a measure we refer to as *Review Karma* (RK).

In the following we first discuss related work regarding RK, feedback quality and motivation; we then explain how RK is calculated and used in our online learning system, followed by a description of our efforts to evaluate RK. Finally, we discuss how the introduction of RK directly influenced [blinded name]'s learning design and draw conclusions on how future iterations could work to circumvent its shortcomings.

2. Related Work

In a time of MOOCs, depending on expert reviews is not sustainable given the number of participating students. The resulting drive towards systems based on peer reviewing and evaluation has generated a sizeable body of work on peer feedback and reviewing.

It has been found that peer reviewing in a learning context helps to generate new ideas and insights (Nagel & Kotzé 2010) and eliminates problems with given solutions earlier than in normal evaluation cycles (Garousi 2010), but that its impact is directly affected by quality, reliability and validity (Gielen et al. 2010). Peer reviews are also often biased towards a higher score in comparison to expert reviews (Papadopoulos et al. 2012), (Lu et al. 2015), which might distort a student's perspective on the quality of their work.

The students' attitude towards peer reviewing is discussed in a number of papers. Some describe a positive view on peer review (Bauer et al. 2009) or seeing it as generally useful (Basnet et al. 2010), others report students finding it hardly helpful and are complaining about a lack of effort from their peers (Lu et al. 2015), (Nagel & Kotzé 2010), or receiving mostly offensive feedback from their colleagues (Wolfe 2004). Further issues mentioned in related research show some of the reservations students have against reviewing, e.g. that reviewing is teachers' work that makes students uncomfortable (Basnet et al. 2010), that some criteria were hard to judge (Bauer et al. 2009), that reviewing took time away from other work (Basnet et al. 2010), and that reviewing in groups showed instances of social loafing (Turner et al. 2011).

There is some discussion on how to motivate students to write good reviews by catering to intrinsic and extrinsic motivation. Extrinsic motivation can be applied to keep students engaged in the reviewing process (Turner et al. 2011), e.g. an influence on grades (Joyner 2016) or open access to assignment solutions (Neubaum et al. 2014). Some work focussed on feedback quality and its influencing factors. A number of papers discuss different influences of qualitative (e.g. text-based) and quantitative (e.g. numeric or categorized) feedback in reviewing (Hicks et al. 2016), (Kulkarni et al. 2015a), (Kulkarni et al. 2015b).

There is a lot of work geared at optimising the peer review process, be it organisational issues in large courses (e.g. Papadopoulos et al. 2012), quality enhancement (e.g. Hicks et al. 2016) or student motivation (e.g. Turner et al. 2011). However, seeing that peer reviews become increasingly more important in today's educational landscape, we need ways to ensure a certain quality over the whole duration of a course, find ways to correctly identify helpful, average, meaningless or offensive reviews as is also tried in reviews in the context of e-commerce (Kim et al. 2006), and to adequately react to a rise or drop of review quality. Towards this goal, we are looking into using learning analytics in the form of a RK, which are in turn directly influencing our overall learning design.

3. Review Karma

Review karma is a term we introduced to describe a value designed to indicate the overall quality of reviews a student (see Luckner & Purgathofer 2015). In our system RK is calculated for each student and lecture, so that RK can reflect the varying interest of students in the respective lectures' subjects, influencing their motivation, willingness or capacity to provide good reviews.

RK is based on relevant information about a student's reviewing ability. It is calculated from feedback reviewees give to reviewers by choosing one of four feedback categories, and from lecture staff flagging especially good or bad reviews. Reviewer skills are hence

judged by other students and lecture staff alike, as is done in other reputation systems where reputation is 'a claim about a target made by a third party' (Farmer & Glass 2010).

Our calculation merges the quality of a student's reviews and their overall participation in the review process to produce a robust metric that represents a student's RK. The calculation results in a relative ranking of students in regard to their reviewing ability.

$$p_s = \textit{set of all positive feedback a student s has received in a lecture}$$
$$n_s = \textit{set of all negative feedback a student s has received in a lecture}$$
$$P = \{p_1, p_2, p_3, \ldots\}, \quad N = \{n_1, n_2, n_3, \ldots\}$$

$$\textit{Review karma } r(s) = \frac{\sum p_{si} * w(p_{si})}{\sum p_{si} * w(p_{si}) + \sum n_{si} * w(n_{si})} * \frac{|p_s| + |n_s|}{|P| + |N|}$$

The first multiplier calculates the relative value of positive feedback received for reviews of a student *s*. It is defined by calculating the percentage of weighted positive feedback in relation to the total weighted feedback, the sum of weighted positive and weighted negative feedback. The weights *w* assigned to types of feedback were initially chosen based on existing karma algorithms such as used in Movshovitz-Attias et al. (2013). We iteratively analysed the results of the calculation, looking for outliers that were not classified fittingly. Now, feedback is weighted according to optimized values that reflect reliability and significance of the source it came from. The second part of the RK calculation introduces how much the student *s* was involved in the review process of the lecture as of yet. The result of this formula is a value that gives a relative ranking of student feedback quality. Finally, experienced evaluators were consulted to sense check the results of the formula, resulting in a last adjustment of the weights used for the RK calculation. The evaluators were shown a segment of students' data the algorithm had ordered into the same category along with all information used to calculate the RK. They were asked to indicate, which category they would place these students in, which was then compared with RK results.

4. Survey

Students from two lectures were invited to participate in a survey and offered points towards their final grade. 146 students completed the survey. It was evaluated based on the qualitative content analysis method described by Meyring (2003). The online survey consisted of 15 open-ended questions. The focus was to collect opinions and experiences with peer reviewing and the task system used in the lecture [blinded ref] and to grasp the impact and effect of the RK or measures used to calculate RK. Results of this survey will be discussed in the next paragraphs.

Reading reviews led to different learning outcomes. Additional to new insights concerning

the lecture content, students also learned about giving helpful feedback and gained motivation from well-written reviews. Badly written reviews, however, had a negative influence on the motivation and led to a feeling of frustration. Only 11.2% of the students indicated that they did not learn anything from reviews they received.

Writing reviews was also mostly perceived as a positive influence on their learning (80.4%) but some people felt their work being actively interrupted by the review system (6.3%) and saw no use in putting any effort into reviewing. Students observed being negatively affected by repeatedly having to review bad work of their colleagues. Seeing bad work handed in by their peers also seemed to devalue their own effort. Most students (77.6%) were able to draw from reviews they received as best practice examples. Especially students with no prior experience in peer reviewing noticed feedback skill improvements. Students with previous knowledge experienced loss of timidity, learned how to write less offensive feedback and be constructive. Others mentioned becoming more motivated to write good reviews over time, especially when receiving 'helpful' as feedback for a critical review. For some students, writing reviews was more fruitful for their own learning than receiving reviews, and it helped them improve their own work by better understanding the requirements or by gaining insight from their peers' perspectives. By proposing changes and critiquing their peers' work they also found similar aspects lacking in their own work.

Students commented on the RK and how it affected their ability to give feedback. While RK was seen as a motivation to improve reviewing skills, others thought it a way to punish critical reviews. Regarding the feedback they received for their reviews, many students expressed scepticism that their peers could differentiate between a good but critical review and a badly written or spiteful review. Some students also mentioned that they felt bad when giving deserved negative feedback to reviews as that could have a negative effect on reviewers' grades. Such influences on the review feedback behaviour create problems since meaningful calculation of RK is dependent on honest reviewee feedback.

5. Discussion

As indicated by the survey results, students experience reviewing and being reviewed overall as a positive impact on their learning, especially when they accept reviewing as an essential part of their work in the course. This positively perceived impact stems from writing reviews to high quality work and receiving well-written reviews to their own.

Students mentioned two distinct advantages of writing reviews: being enabled to double-checking their own work and seeing different solutions to a task broadens their horizon, as also observed by Nagel and Kotzé (2010). By getting students in the habit of critiquing other people's work, they also reflect on their own work and learning. Some students even

went so far as to value writing reviews over receiving reviews, saying they learned more from analysing their peers' work than getting feedback on their own work.

While RK is a value calculated from a large amount of data emanating from the system, most of this data is a direct reflection on how reviews are received by the reviewers' peers and the lecture staff. Answers to the survey, however, show that this social component of the RK introduces a dilemma for students writing reviews: students hesitate to write critical reviews because they fear it being negatively perceived and hence down-voted with negative review feedback. However, the survey also shows that students long to receive critical feedback on their work because it is perceived as more helpful. This shows a conflict between their own experience, that receiving a critical review helps them more, and their behaviour, which is based on the notion that writing a critical review will get them negative review feedback that in turn influences their grade. It leaves us in a difficult spot between fostering critical reviews and basing their grades on reviews they write and the feedback they get from their peers. While we are not aware of this particular dilemma being discussed in literature as of yet, there are some comments about the need to deal with offensive feedback in Wolfe (2004). Lu et al. (2015) raised the issue of unhelpful feedback being particularly bad for at-risk learners, underlining the importance of this dilemma.

Reading well-written reviews that prompted students to revise their works was often seen as educational beyond the suggested changes to and enhancements of the original work. Such reviews acted as samples of how to write helpful feedback, directly impacting the peer learning process. Some students started to emulate best practices they found in reviews they received, describing those practices as a positive learning outcome, which confirms Lu et al.'s (2015) notion that feedback gets better if reviewers put in more work. While students did not hesitate to give good review feedback to reviews they liked, they struggled to indicate which reviews they found unhelpful or meaningless. Students mentioned having a bad conscience when negatively evaluating their peers' reviews and would rather choose 'average' as review feedback than truthfully evaluating these reviews negatively. It follows that students are troubled voicing their real opinion about their peers' work as they get the feeling that they would create trouble for their peers or influence their grades, supporting Basnet et al.'s (2010) observation that students do not like the idea of marking their peers.

Since we learned that students emulate best practices from reviews they received, it stands to reason that receiving more well-written reviews might even help write better reviews. Hence, we introduced targeted allocation, as also suggested by Kulkarni et al. (2015). Students with high RK are now assigned the work of students with low RK for reviewing. If receiving well-written reviews increases the quality of the reviews a student writes, we hypothesise that an underlying self-enhancing system could be put at work here. Targeted allocation of reviews can be risky from an ethical viewpoint, as students with high and average RK would receive fewer reviews from students with good RK. However, our basic

assumption is that targeted allocation would raise the review quality throughout the course, which in turn would benefit everybody. Assuming this is a viable model, we only need to create a critical mass of good reviews in order to bring the system into a state of positive self-enhancement. In this sense, we see the initial detraction of high-quality reviews from average students as an investment that ultimately benefits everybody.

Another change in the design is to put a higher value on reviewing by increasing its influence on the final grade. This correlates with the results of the survey showing that review writing has a large impact on students' learning progress, as is supported by Joyner et al. (2016). Such a measure can enhance the perceived importance and incite students to spend more time and effort in writing reviews. An extension of this would be to make RK of reviewers visible to reviewees, helping students to contextualise reviews they receive.

6. Conclusion and Future Research

Especially in large classes, it is hard to keep an overview of individual students' work, and to intervene when needed, so that we do not leave struggling students behind. RK is one of the possible tools to support the lecture team in this. We are still striving to iron out the kinks in the current RK implementation, and the evaluation of our design changes is pending. Our experience as well as the students' observations are pointing towards some interesting challenges for future design iterations such as how to make reading reviews more interesting; how to raise the overall review quality; how to get students to provide honest review feedback instead of defaulting to 'average' for meaningless or offensive reviews; or how to foster a climate that supports writing critical and helpful feedback that does not inadvertently trigger a trend of writing solely positive reviews in fear of a review feedback backlash. These and further questions will be the focus of our future work in the hope of building a sustainable reviewing environment to support students in their learning.

References

Basnet, B., Brodie, L., & Worden, J. (2010). Peer assessment of assignment. In *Frontiers in Education Conference* (FIE), 2010 IEEE (p. T1G–1–T1G–2).

Bauer, C., Figl, K., Derntl, M., Beran, P. P., & Kabicher, S. (2009). The Student View on Online Peer Reviews. Proceedings of the 14th Annual ACM SIGCSE Conference on Innovation and Technology in Computer Science Education., 41(3), 26–30.

Farmer, R., & Glass, B. (2010). Building Web Reputation Systems (1st ed.). USA: *Yahoo! Press*.

Garousi, V. (2010). Applying peer reviews in software engineering education: An experiment and lessons learned. IEEE *Transactions on Education*, 53(2), 182–193.

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction*, 20(4), 304–315.

Hicks, C. M., Pandey, V., Fraser, C. A., & Klemmer, S. (2016). Framing Feedback: Choosing Review Environment Features that Support High Quality Peer Assessment (pp. 458–469). Presented at the Proceedings of the 2016 *CHI Conference on Human Factors in Computing Systems*, ACM Press.

Joyner, D. A., Smiley, A., Bruckman, A., Goel, A., Ashby, W., Irish, L., … Sheahen, D. (2016). Graders as Meta-Reviewers: Simultaneously Scaling and Improving Expert Evaluation for Large Online Classrooms (pp. 399–408). Presented at the Proceedings of the Third (2016) ACM *Conference on Learning @ Scale*, ACM Press.

Kim, S.-M., Pantel, P., Chklovski, T., & Pennacchiotti, M. (2006). Automatically assessing review helpfulness. In Proceedings of the 2006 *Conference on empirical methods in natural language processing* (pp. 423–430). Association for Computational Linguistics.

Kulkarni, C. E., Bernstein, M. S., & Klemmer, S. R. (2015). PeerStudio: Rapid Peer Feedback Emphasizes Revision and Improves Performance (pp. 75–84). Presented at the Proceedings of the Second (2015) ACM *Conference on Learning @ Scale*.

Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., … Klemmer, S. R. (2015). Peer and self assessment in massive online classes. In *Design Thinking Research* (pp. 131–168). Springer.

Lu, Y., Warren, J., Jermaine, C., Chaudhuri, S., & Rixner, S. (2015). Grading the Graders: Motivating Peer Graders in a MOOC. Proceedings of the *24th International Conference on World Wide Web*. 680–690.

Luckner, N., & Purgathofer, P. (2015). Exploring the Use of Peer Review in Large University Courses. *Interaction Design and Architecture(s) Journal*, (25), 21–38.

Mayring, P.: Qualitative Inhaltsanalyse. *Grundlagen u. Techniken.* Weinheim: Beltz (2003)

Movshovitz-Attias, D., Movshovitz-Attias, Y., Steenkiste, P., & Faloutsos, C. (2013). Analysis of the reputation system and user contributions on a question answering website: Stackoverflow. In *Advances in Social Networks Analysis and Mining*, 2013 (pp. 886–893).

Nagel, L., & Kotzé, T. G. (2010). Supersizing e-learning : What a CoI survey reveals about teaching presence in a large online class. *The Internet and Higher Education*, 13(1–2), 45–51.

Neubaum, G., Wichmann, A., Eimler, S. C., & Krämer, N. C. (2014). Investigating Incentives for Students to Provide Peer Feedback in a Semi-Open Online Course: An Experimental Study (pp. 1–7). *ACM Press*.

Turner, S., Pérez-Quiñones, M. A., Edwards, S., & Chase, J. (2011a). Student Attitudes and Motivation for Peer Review in CS2. In Proceedings of the *42Nd ACM Technical Symposium on Computer Science Education* (pp. 347–352).

Papadopoulos, P. M., Lagkas, T. D., & Demetriadis, S. N. (2012). How to improve the peer review method: Free-selection vs. assigned-pair protocol evaluated in a computer networking course. *Computers and Education*, *59*(2), 182–195.

Wolfe, W. J. (2004). Online Student Peer Reviews. In Proceedings of the *5th Conference on Information Technology Education* (pp. 33–37). New York, NY, USA: ACM.