Peer Review Data Warehouse: Insights from Different Systems

Ferry Pramudianto[1], Maryam Aljeshi[2], Hisham Alhussein[3], Yang Song[4], Edward F Gehringer[5], Dmytro Babik[5], David Tinnaple[6]

[1,2,4,5]Department of Computer Science, North Carolina State University, NC
[3]Department of Computer Science, University of Massachusetts Amherst, MA
[5]College of Business, James Madison University, Harrisonburg, VA
[6]Herberger Institute for Design and the Arts, Arizona State University, AZ
[[1]fferry, [2]mmaljesh, [4]ysong8, [5]efg]@ncsu.edu, [3]hisham.hussain@kaust.edu.sa, [5]babikdm@gmail.com,
[6]david.tinapple@asu.edu

ABSTRACT

Peer assessment is widely used at all levels of education. Students give and receive feedback from their classmates, and thereby produce a wealth of information that can potentially be used to improve the assessment process. But thus far, each online peer- assessment system has been an entity unto itself. There has been no attempt to compare the approaches taken by such systems, for example, the rubrics or the structuring of the assessment process. Our PeerLogic project is an attempt to change that. We are constructing a data warehouse of millions of peer reviews, from at least half-a-dozen systems, that can be mined to determine how differences in the assessment processes translate into differences in peer assessments. This paper reports on some of the issues that arise in the construction of the warehouse, and how we have resolved them in a way that will work for all constituent systems. We also presented an example of comparing data coming from two systems that are based on rating and ranking.

Keywords

Peer review, data warehouse, data modeling, data mining, peer-assessment.

1. INTRODUCTION

In recent years, many papers have been published on individual peer-assessment/review systems [1-4]. But invariably, the data--and the conclusions--are all derived from a single system. There is currently no easy way for the educational peer-review research community to share their data. One hurdle is that these works sometimes use different terminology for describing the same things. For instance, the work that is to be peer assessed may be called a "submission," an "artifact," or an "answer." The assessment given by the peer can be referred to as a "review," a "critique," or "feedback."

In addition, peer-review systems were developed based on different design choices. For example, some systems let a reviewer rate the artifact on a Likert scale (or multiple Likert scales for several criteria). Other systems ask reviewers to rank the artifacts against each other. For another example, some systems structure an assignment as a set of submission and review tasks. Other systems handle these tasks as different assignments. Someone who is trying to combine the data from multiple systems needs a thorough understanding of how these systems work, and this will take much time to achieve.

With these differences, it is fairly difficult for researchers to share data and perform comparison studies [1]. But this kind of research is important, because only through it can we determine which of the different design choices are most effective in promoting learning gains. Toward that end, we present a Peer-Review Markup Language (PRML), which defines markup for modeling metadata for peer-review activities. PRML is designed to be a generic data model/schema for modeling and sharing numeric and textual data from multiple peer-review platforms/applications. PRML was designed jointly by the originators of four online peer-review systems, Expertiza [2], Mobius SLIP [3], CritViz [4], and Crowdgrader [5]. It is intended to define a common terminology for the concepts used in educational peer review, and to allow a system designer, researcher, or practitioner from any educational domain to use the same vocabulary when talking about peer assessment. Secondly, by using its common set of concepts, PRML can provide an overview of how online peer assessment works in practice. Thirdly, PRML can serve as a foundation for creating a shareable "data warehouse" that can be used by any peer-assessment researcher. The sheer number of reviews allows them to study and compare the effects of different peer review approaches with a stronger statistical power.

This paper is structured as follows. In section 2 we describe the PRML design. Section 3 gives the design and implementation of the data warehouse, and in Section 4 we present an example of a data analysis. Section 5 concludes our paper and suggests future work.

2. PRML Design

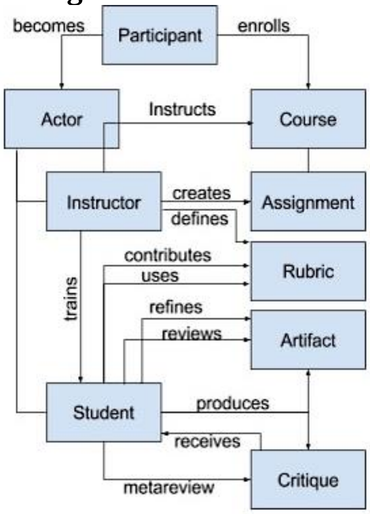

Figure 1. PRML Main Concepts and Their Relationships.

The core of the PRML describes the relationship among entities involved in the peer-review process (figure 1). It includes a Participant, who is enrolled in a course, which has one or more Assignments, and The assignments can be undertaken by individuals or groups. Each individual or group may also work on the same task (e.g., everyone writes the same program), or they may work on different tasks (say, papers on different topics, or different modules for an open-source software application). Within an Assignment, the individuals and groups are abstracted as Actors. Actors can be categorized according to their role within the assignment as Instructors and Students. In practice, some students can also be teaching assistants and play the role of instructors. The instructors instruct the course, create assignments and rubrics for evaluating students' work. A Rubric can either be holistic, or criterion-based. In any case, the holistic and criterion based rubric may contain a particular type of prompt: an open-ended question, a multiple-choice question, a checkbox, or a Likert-style rating, and so forth.

3. DATA WAREHOUSE

We derive a DW model from PRML that can be used to share data from different peer-review systems. It was designed based on dimensional modeling (DM) approach [6]. DM stores measurements, metrics, or facts of the business process in tables, referred as Fact tables that hold references to the dimension tables (foreign keys). The Dimension tables contain groups of hierarchies and descriptors that define the facts. The dimensions can be used to group the facts into multidimensional arrays of data, known as OLAP cube or hypercube. DM encourages DW schema to follow a star topology, in which fact tables are placed in the center.

Following this approach, our schema is centered around the Critique table since it contains measurements to the artifacts that are expressed through the reviewer's qualitative and quantitative feedback. The quantitative feedback can be expressed in rating, ranking or the combination of both. This design choice allows us to group the feedback according to various dimension e.g., student performance in particular topics, assignments, and courses, as well as comparing the effect of various peer assessment approaches to student's performance.

As depicted in Figure 2, the Critiques can be sliced based on different dimensions including the Criterion, Eval_Mode, Task, Actor, and Course_Setting. The criterion table contains criteria questions, the scale used to rank or rate the work, and the weighting that is used to calculate the final score. The Eval_Mode determines whether ranking, rating or both are used to evaluate the artifact. The Task table contains information such as when the task starts and ends, the CIP (Classification of Instructional Programs) codes, whether it is an assignment, reviewing, or meta-reviewing task. The actor table contains the actors involved in the assignment and their roles, whether it is a student, instructor, or administrator. The actor table is linked to the participant table in the Actor_Participant table to maintain the group memberships of each participant. The Artifact table contains information about the student's work in response to the assignments, which can be stored as a plain text or URLs to uploaded files and web pages. The Course_Setting table contains meta-data about how the peer review was conducted that can be used to compare the effect of different features adopted by peer review systems to the learning gains as well as the quality of the peer review process itself. The examples of the meta-data include for instance: Anonymity, which specifies how anonymous the reviews are. Several approaches could be adopted e.g., the authors and reviewers are visible to each other (non-anonymous), the reviewers get to see the authors' name but not the other way around (single blind), The authors and reviewers are anonymized to each other (double blind). Another approach could initially perform reviews anonymously, then after the process are finished,

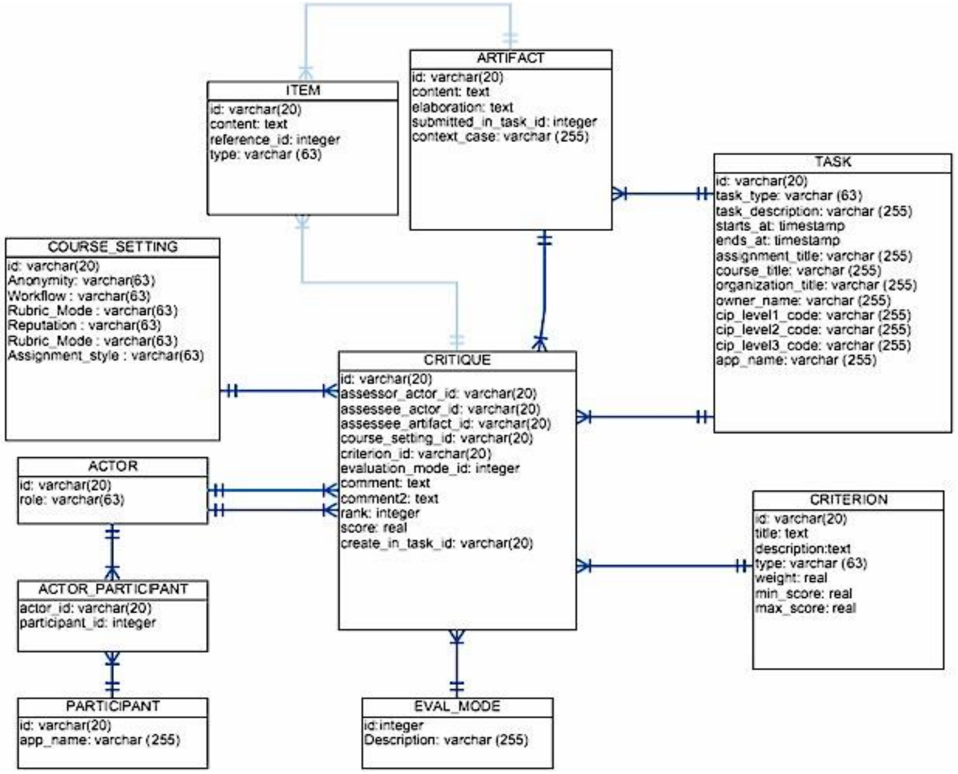

Figure 2. Data Warehouse Schema.

the reviews are de-anonymized (everybody could see who wrote the feedbacks to the artifacts) to provide a sense of accountability.

The Course_Setting table also contains meta-data that shows if students participate in multiple rounds of review for the same artifact. Multiple rounds of reviews may be used to provide unidirectional feedback from the reviewers to the authors, but they could also be used to let the reviewers know how helpful their reviews were (feedback from the reviewers to the authors about their work, then authors provide feedback to the reviewers about the usefulness of the reviews). In addition, the Rubric_Mode column specifies if the reviewers should provide holistic reviews or detailed reviews based on certain criteria. The Assignment_Style denotes if the assignment consists of a fixed set of activities, or if the activities vary from assignment to assignment.

We decided to implement the initial version of the DW using MySQL for several reasons. First, it offers a mainstream query language (SQL) and more mature tools compared to NoSQL databases. Secondly, most peer review systems that we know still use relational databases, therefore mapping them to relational DW would be simpler and less risky than using NoSQL approach. Third, we anticipate, based on the past growth of several systems, that the amount of aggregated data will not exceed 100GB within the next three years, and therefore MySQL would still be able to serve our needs.

During the transformation process, each peer review system runs a Pentaho [7] instance that read their existing database, transform the data according to the DW schema, and load them into a staging DW. After the transformation is validated at the staging DW, another instance of Pentaho populates the data from the staging Data Warehouses into the central DW. This approach is adopted to protect the central DW being corrupted by transformation errors.

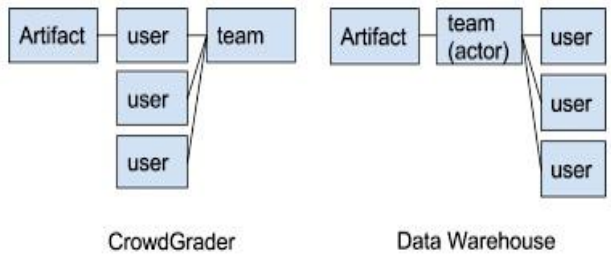

Figure 3. An Example of Changing Entities Relation

Since each system stores peer-review data differently, we need to map their schema to the data warehouse schema. The process was challenging since we needed to combine data within and across tables, split data within the same tables, and introduce new data to preserve the original relationships within the new schema. For instance, Expertiza stores teams and individual actor in separate tables. When we migrate this data into the DW, we have to create a new actor for each individual participant and their mappings in the DW. Another example, CrowdGradder only has a user table. To maintain group assignments, only the leader of the group is linked to the artifact and all members of the team are stored in the team table, which is similar to actor table. When we migrated this to the DW, we needed to replace the link between assignments and teams' leaders to the id of the teams (Figure 3).

In addition, the ETL process also anonymizes the DW by removing any personal information about the students, such as name, emails, and their campus IDs. When the systems store the artifacts as plain text, they can be transferred to the DW and shared. When they are stored as external files or wiki articles, it is up to each system if they allow these files to be shared. The DW only stores the link to these artifacts.

The DW is currently accessible through the MySQL server at PeerLogic.csc.ncsu.edu. We provided a read only credential upon request. In the future, these data will be accessible through a RESTful web service and visualized through our website at PeerLogic.org.

Since the DW receives data from widely used systems, the amount of data is quite extensive in many dimensions: the number of participants, the number of peer reviews, the diversity of peer-review processes, and the range of students' background and level. Researchers can mine this dataset to derive general conclusions rather than conclusions that are just class or task specific. Once we have completed transferring data from the four systems, we will be able to examine our hypotheses using a much larger dataset, including courses in various disciplines, held on several campuses. We can also analyze the qualities of the feedback and explore correlations to the approaches used by different systems.

4. DATA WAREHOUSE USAGE

As an example on how we could use the data warehouse, we conducted a study to compare peer review systems which are based on rating and ranking. In addition, we also present our literature research on these two systems to provide a context how these systems diverse.

Rating and ranking are both used as peer assessment grading scheme and have been proved to be useful by research; It is important to discuss how both tools differ from the student's, reviewer's, and instructor's perspectives. Peer rating evaluates an assignment based on specific scale, while in peer ranking a group of students' works are ranked from best to worst. Students will most likely have different reactions based on how well they do in both of these tools: Good ratings can motivate a student by receiving high ratings in different criteria. It also shows their strengths. Good rankings, however, may motivate a student by reassuring them they were graded highly amongst others in the class, and that student will either be less satisfied knowing that their colleagues ranked even higher than them or see this as an incentive to get a better rank next time. In contrast, bad ratings may demotivate a student when compared to the full rating score for the assignment, but the obscurity of their standing in the overall class and colleagues' performances may provide some relief in the sense that there may be others who had worse ratings. On the other hand, bad rankings can demotivate students by illustrating their worst performance in comparison with their cohorts. When ranking is used among best performers, it could be very frustrating since it raises the competitiveness among the best students, but on the other hand, it could also motivate them to perform even better and also train them to face the real world situations beyond their academic life, where competitions are inevitable. In the contrary, when it is used amongst low performer students, it could lead to premature satisfactions, which does not help them to reach their full potential.

From the reviewer's angle, rating provides more flexibility and accuracy opposed to ranking; for instance, two students that have excellent papers can both have a full rating, but in ranking further observation needs to be done to decide which one should be ranked higher. That being said, ranking can be more time-consuming for the reviewer and less accurate than rating is.

Table 1. Comparison Between Ranking and Rating

	Peer Rating	Peer Ranking
Definition	"assessment of each member, by the rest of the group, on a set of performance characteristics" [7].	"ranking all individual group members from best to worst against a given set of characteristics" [7].
Accuracy	A research by Freeman & Parks comparing scores assigned by students and scores by experts to question if students grading was good enough to use. It was addressed in the paper that "This question is subjective and context-dependent; in our case, the answer is yes." [10]	It is difficult to judge the quality of someone's work using ranking since it only shows how it compares to the others. Similar quality works might get different ranks.
Usefulness	Increase the confidence of the grades if the reviewers agree on approximately similar grades.	Tinapple et al. [4] state that "CritViz quantitative ranking system we have is not designed as a grading mechanism, but rather for self-organization and reflection as a class".
Advantages	Pope mentioned in his article that some studies that were successful in using peer rating concentrated in few disciplines such as law, business, and nursing. Some of the advantages were [8]: • improvement in final papers • reinforcement of learning and exchange of ideas • understanding of marking schemes and standards • understanding of presentation skills • change of the teacher's role to facilitator from assessor	Tinapple et al. [4] discuss the use of CritViz which is a web software that uses peer critique-- peer ranking in a large classroom. Students reported feedback about the system: • It enhanced the feeling of community in the classroom • Allows students to share ideas
Disadvantages	Prone against collusion, where students rate each other quite high regardless the actual quality of the work. Rating is susceptible to individual biases since everyone has different standards.	"there may be a variable number of rankings for each item, which can lead to the scoring of certain items being more reliable than others" [9] "Ranking too many items might easily lead to sloppy ranking and, thus, bad data" [9].

However, rating could be abused by a group of people who conspire to give each other good ratings. Lastly, an instructor incorporating the rating system will have a better indicator of how well the students' mastery of a topic; since it provides ratings based on different criteria. On the other hand, an instructor that incorporates the ranking system will only have a distribution of the students on a spectrum, which does not say much about their competence in different criteria, and thus is a bad indicator of student mastery. Again, the rating system is less time consuming for the instructor since it groups feedback based on specific criteria.

A possibility to compare the effect of rating and ranking is through measuring the quality of the feedback in these different systems. The quality can be examined through the amount of the feedback, the type of the feedback (e.g., problem detections, praise, improvement suggestions), or the tone polarity.

4.1 Insights from Data Warehouse

As an initial attempt to mine information out of our data warehouse, we try to compare the feedback volume of two different systems that use ranking (Expertiza) and ratings (CritViz). By understanding the DW schema, we were able to easily query the data and found that in average the feedback in Expertiza is 329 words long (SD=267), and 184 (SD=140) as depicted in Figure 4. This simple information would have been quite difficult to obtain when we have to deal with different systems and database schemas. Since we would have to understand in detail how CritViz and Expertiza store this information in their database in order to design the SQL scripts to mine the information.

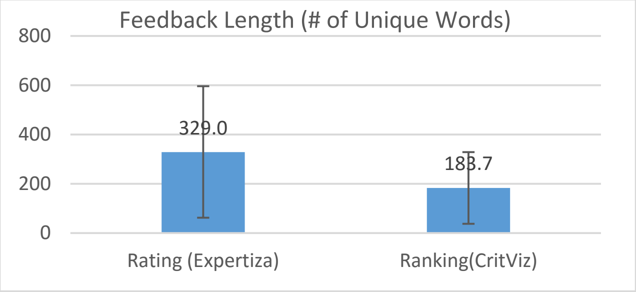

Figure 4. The average of feedback unique words in Expertiza and CritViz

However, we would like to stress that this simple comparison cannot be used to generalize the different between ranking and rating since we believe that the volumes are also influenced by how the rubric is designed. In Expertiza, the instructors usually design a rubric with multiple criteria, upon which the submissions should be rated by the reviewers. In addition, the reviewers should provide a qualitative feedback on each criterion to justify their ratings. There is no limit on how many criteria that a rubric should contain. But on average the rubrics in Expertiza contains 5-8 questions, while in CritViz average between 4-5.

The number of criteria is not the only factor which prompts users to give extensive feedback, but also the creativity of the reviewers and their motivations play major roles. In addition, the type of questions being used as criteria also plays an important role in triggering the reviewers giving useful feedback. For instance, short answer questions will likely prompt less extensive, but more consistent feedback. Meanwhile, open-ended questions could lead to more fruitful feedback but hard to quantify.

5. CONCLUSION AND FUTURE WORK

We use these common terms in the PeerLogic project, involving four different systems, dealing with diverse disciplines such as computer science, business, art, and education. Although we have not yet performed scientific studies to evaluate our approach, the project members communicate constantly using these common terms. They agree that having a common dataset and terminology help simplify the collaboration in peer review community. Although they deal with different domains, they are able to understand each other when talking about the peer assessment concepts. Having a common DW also helps the researcher to share their data and compare them with the results other peer review studies.

At the moment, we only support transforming data through ETL tools. In the future, we would like to provide a web interface that allows instructors to share their data simply by uploading a comma separated value (CSV) files. We would also provide a user interface to help visualize these data. Moreover, we plan to provide a set of common web services that help researcher run comparable future studies using different systems. These web services will include visualization, meta-review, reputation, and reviewer assignment.

We would also like to mine more information out of the data warehouse to compare different properties of the systems e.g., holistic vs detailed rubric, other effects of rating and ranking, multiple vs single round, as well as different visualization techniques.

6. ACKNOWLEDGMENTS

This study is partially funded by the PeerLogic project under the National Science Foundation grants 1432347, 1431856, 1432580, 1432690, and 1431975

7. REFERENCES

[1] Luxton-Reilly, A.: ‘A systematic review of tools that support peer assessment’, Computer Science Education, 2009, 19, (4), pp. 209-232.

[2] Gehringer, E., Ehresman, L., Conger, S.G., and Wagle, P.: ‘Reusable learning objects through peer review: The Expertiza approach’, Innovate: Journal of Online Education, 2007, 3, (5), pp. 4.

[3] Babik, D., Singh, R., Zhao, X., & Ford, E. (2015). What you think and what I think? Studying intersubjectivity in evaluation of knowledge artifacts. Information Systems Frontiers, 1-26. DOI: 10.1007/s10796-015-9586-x.

[4] Tinapple, D., Olson, L., and Sadauskas, J.: ‘CritViz: Web-based software supporting peer critique in large creative classrooms’, Bulletin of the IEEE Technical Committee on Learning Technology, 2013, 15, (1), pp. 29.

[5] de Alfaro, L., and Shavlovsky, M.: ‘CrowdGrader: A tool for crowdsourcing the evaluation of homework assignments’, in Editor (Ed.)^(Eds.): ‘Book CrowdGrader: A tool for crowdsourcing the evaluation of homework assignments’ (ACM, 2014, edn.), pp. 415-420.

[6] Kimball, R., and Ross, M.: ‘The data warehouse toolkit: the complete guide to dimensional modeling’ (John Wiley & Sons, 2011. 2011).

[7] Pulvirenti, A. S. (2011). Pentaho Data Integration 4 Cookbook. Packt Publishing Ltd.

[8] Pope, N.: ‘An examination of the use of peer rating for formative assessment in the context of the theory of consumption values’, Assessment & Evaluation in Higher Education, 2001, 26, (3), pp. 235-246.

[9] Waters, A.E., Tinapple, D., and Baraniuk, R.G.: ‘BayesRank:A Bayesian Approach to Ranked Peer Grading’. Proc.Proceedings of the Second (2015) ACM Conference on Learning @ Scale, Vancouver, BC, Canada 2015.

[10] Freeman, S., and Parks, J.W.: ‘How accurate is peer grading?’, CBE-Life Sciences Education, 2010, 9, (4), pp. 482-488