UtahStateUniversity
EMMA ECCLES JONES COLLEGE OF EDUCATION & HUMAN SERVICES
INSTRUCTIONAL TECHNOLOGY & LEARNING SCIENCES

Modeling Teacher Ratings of Online Resources: A Human-Machine Approach to Quality

Mimi Recker, Heather Leary, Andrew Walker, Anne Diekema
Utah State University
2830 Old Main Hill
Logan, Utah 84322
1.435.797.2692
heatherleary@gmail.com

Philipp Wetzler, Tamara Sumner, James Martin
University of Colorado
594 UCB
Boulder, CO, USA
1.303.735.4469
philipp.wetzler@colorado.edu

Paper presented at the American Educational Research Association Conference,
New Orleans, LA
April 2011

OBJECTIVES

The increased pervasiveness of networked computing coupled with a participatory web culture has spawned new models of innovation and creation. These models, variously referred to as collective intelligence, crowd sourcing, wisdom of crowds, collective intelligence, or peer production, occasionally take the Internet by storm (Benkler, 2006). As canonical examples, Wikipedia and YouTube need no introduction.

Similarly, in education, the scalable deployment of media-rich online resources supports peer production in ways that promise to radically transform teaching and learning (CRA, 2005; Pea et al., 2008). For example, a growing trend toward sharing online instructional resources has spawned the global *OpenCourseware* movement (Smith & Casserly, 2006). Likewise, online educational repositories such as the Digital Library for Earth Systems Education (DLESE.org) and the National Science Digital Library (NSDL.org) collect and curate online learning resources created for a wide range of educational audiences and subject areas (McArthur & Zia, 2008).

We have developed a simple, web-based authoring tool, called the Instructional Architect (IA.usu.edu), which enables teachers to freely find, gather, and produce instructional activities for their students using online learning resources (Recker et al.,

2005). Teachers can share these resulting activities, called *IA projects,* by making them publically available on the Web. These IA projects can then be viewed, copied, and adapted by other IA users, in ways that support innovative teacher peer production.

A vexing problem for such initiatives remains the elusive notion of *quality.* In peer production environments, how does one identify quality online content? Moreover, how does one do so in sustainable, cost-effective, and scalable ways? Previous work (Bethard, et al, 2009) presented an innovative approach for using machine learning models to automatically assess the quality and pedagogic utility of educational digital library resources. They demonstrated the feasibility and accuracy of automatic quality assessments for a single STEM domain and audience-level: high school Earth science. In this article, we report on recent efforts to extend these models to support a broader range of STEM topics and grade levels. Specifically, we applied the quality models to 200 IA projects and compared model outputs to quality assessments made by K-12 teachers. Since the nature of the resources being compared in the IA (peer) versus DLESE (expert) are different, results of this study provide insights on the generalizability of this machine learning approach and its potential for facilitating teacher peer production.

THEORETICAL FRAMEWORK

Approaches to Quality in Existing Educational Digital Libraries

Many digital library builders have established review rubrics to evaluate the quality of online educational resources generated by users, including teachers and faculty (e.g., Fitzgerald, Lovin, Branch, 2003; Lamantia, 2003; Maurer & Warfel, 2004). Some digital libraries have developed rubrics to identify high quality online resources in order to establish their reputation with users (Sumner, Khoo, Recker, Marlino, 2003), while others use rubrics to determine the inclusion of high quality resources in a digital library or repository (Lamantia, 2003; Maurer & Warfel, 2004). One site also developed a rubric to guide authors in creating high quality online resources (McMartin, 2004) by gathering feedback from target users. All of these approaches have tradeoffs; none has shown to be easily implemented, especially at scale (Bethard et al., 2009).

TABLE 1
Quality indicators (Bethard et al., 2009)

Table 1. Quality indicators (Bethard et al., 2009)

Quality Indicators	Baseline Performance	ML Model Performance
Not inappropriate for age	99%	99%
Indicates age range	79%	87%
Has instructions	61%	78%
Identifies learning goals	72%	81%
Organized for goals	75%	83%
Has prestigious sponsor	70%	81%

OPERA Algorithm

The Open Educational Resource Assessments (OPERA) algorithm used in this paper was initially developed for assessing quality in online resources collected in the Digital Library for Earth Systems Education (DLESE.org) (Bethard et al., 2009). Acknowledging that quality is contextual, this approach avoids a unary "thumbs-up/thumbs-down" assessment; instead it relies on characterizing quality as decomposable into a number of indicators with are applied in turn. Based on detailed analyses of teachers' and expert catalogers' rating of resources, a number of salient indicators were distilled (see Table 1).

Each indicator is modeled using a support vector machine, a type of supervised machine learning algorithm. Supervised machine learning algorithms construct models by statistically analyzing a training corpus for which the correct judgment is known. The OPERA models were trained on a set of human-tagged DLESE resources that indicated the presence or absence of each indicator. Through training, the models attempt to learn which features of an online resource help to determine the presence or absence of an indicator. While the support vector machine algorithm has been shown to be effective at detecting relevant statistical patterns even when the number of features is extremely large, the way in which those features are presented to the algorithm greatly influences the resulting model. The initial set of features used in this study was guided by a large corpus of prior work in using machine learning on linguistic and semantic tasks: features used include 'bag-of-words', term frequency, resource URLs, Google page rank, etc.

Evaluation took place using a different set of DLESE resources from the training corpus. Each model's output was compared to the ratings of two expert DLESE catalogers who were asked to judge the presence or absence of the quality indicators in each resource in the evaluation set. Table 1 shows six of the quality indicators used and the percent of time the models agreed with the human experts. Model results are also compared to a simple baseline that always assumes the most common case. For example, the "has instructions" indicator is present in 39% of resources. If we always assume that a resource has no instructions, we'd be correct in 61% of cases. Good improvements over the baseline were achieved on the "has instructions" and "has prestigious sponsor" indicators, and moderate improvements on the "indicates age range" and "organized for goals" indicators.

Study Context: The Instructional Architect

The context for this work is a free, web-based tool, *the Instructional Architect (IA.usu.edu),* used by teachers to author instructional activities for students using online resources. Teachers can use the IA in several ways: the *'My Resources'* area allows teachers to directly search for and save online learning resources from the Web, including Web 2.0 technologies like RSS feeds and podcasts. In the *'My Projects'* area, teachers can select online resources, then sequence and annotate them with text to create learning activities (called IA projects) for their students. Finally, teachers can *'Publish'* IA projects for their own students, or anyone on the Web. These public IA projects can then be viewed or copied by other IA users. It is these key services of collecting, creating, and sharing that support peer production in the IA community (Recker et al., 2005).

Currently, the IA has over 5,200 registered users who have created over 11,000 IA projects using over 50,000 online resources. Figure 1 is an IA screen shot, showing a teacher-created learning activity and an embedded online resource.

RESEARCH DESIGN AND METHODS

The purpose of this study was to formulate an approach for identifying high quality IA projects, using a set of quality indicators based on Bethard et al. (2009) and the OPERA algorithm. Three teachers used the set of quality indicators to rate IA projects, and their ratings were compared to OPERA output to evaluate the algorithm's effectiveness in identifying quality IA projects. The following research questions guided this study:

1. How do teacher IA project ratings compare with each other?
2. How do teacher IA project ratings compare with OPERA?
3. To what extent can the OPERA algorithm be used to identify 'quality' IA projects?

Participants and Procedures

Three science teachers who have used the IA and participated in a teacher professional development workshop participated in the study. They were asked to individually rate 200 IA projects. This was done through the use of a side bar add-on in Mozilla Firefox, in which they were presented with the IA project on the right side of the screen and the six indicators on the left. The teachers clicked radio buttons under the indicators to select their choice (1-5 scale from "strongly disagree" to "strongly agree") for each indicator.

The 200 IA projects selected for rating 1) were publicly available, 2) were viewed at least 20 times, 3) had more than 700 words, and 4) used at least 3 online resources. Six quality indicators were selected (see Table 2) from the initial seven reported in Bethard et al. (2009).

TABLE 2.
Six indicators for rating IA projects.

Quality Indicator	Definition
Has instructions	Tells user how to navigate and use the project
Links to prestigious sponsor	Links to 'prestigious' source or site where the manager or organizer is highly respected
Identifies learning goals	Identifies learning goals and articulates the knowledge/skills a student is expected to acquire
Organized for learning goals	Organizes content appropriately for its learning goals
Identifies age range	Identifies its target student age range by stating the expected age or grade
Content seems appropriate for age range	Provides reading or activities that are neither too difficult nor easy for the given grade level.

OPERA Algorithm

The algorithm previously developed by Bethard et al. (2009) was used to classify the same 200 IA projects along 6 quality indicators. Two versions of the OPERA algorithm were used. First, 'out-of-the-box' OPERA was the same as used in previous research (Bethard et al., 2009). Second, trained OPERA was a version trained on ½ of the IA projects (100 projects), and then asked to classify the remaining 100 projects.

OPERA provides two outputs for each quality indicator on each IA project: a yes/no classification as to whether the IA project possesses the indicator, and a score ranging from 0 to 1, indicating the confidence OPERA has in the yes/no response. This score can be viewed as a pseudo-probability, in that the closer to 0 the score is, the more confidence that the answer is "no", and the closer to 1, the more confidence that the answer is "yes".

Measuring Agreement

Machine learning research has typically relied on the *kappa* statistic to measure agreement, and is robust for nominal data. However, *intraclass correlation* (ICC) and *Krippendorf's alpha* (KA) are both more appropriate for interval data. An ICC statistic reports single and average measures, where the single measure takes into account the object being rated while the average measure accounts for both the object and the difference between raters. The measure in this study used an interval scale (from one to five) applied by three teachers. To account for rater and object variability, we therefore chose to use the ICC statistic

RESULTS

Research Question 1: *How do teacher IA project ratings compare with each other?*

Table 3 shows the distribution of ratings for each teacher. They suggest a possible ceiling effect as each of the teacher's ratings have a negative skew.

TABLE 3.
Descriptive statistics for each teacher rater.

	Minimum	Maximum	Mean	SD
Teacher 1	1	5	4.05	1.53
Teacher 2	1	5	4.06	1.14
Teacher 3	1	5	4.07	1.24

To measure teacher agreement, an intra-class correlation (ICC) between the three teachers' ratings was computed for each indicator (see Table 4; 0.40 to 0.59 represents moderate inter-rater reliability, 0.60 to 0.79 substantial, and 0.80+ outstanding). The overall ICC was .622, a substantial level of inter-rater reliability. All indicators achieved at least *moderate* levels of agreement, except for "links to prestigious sponsor". Two indicators, "content seems appropriate for age range" and "identifies age range", achieved *substantial* levels of agreement. Interestingly, the rank order of the teachers'

ICC on indicators is identical to the agreement between experts used in the previous study (see Table 1).

TABLE 4.
ICC value for teacher ratings on each indicator

Quality Indicator	**ICC for teachers' rating of IA projects (rank order)**
Content seems appropriate for age range	.717
Identifies age range	.677
Has instructions	.566
Identifies learning goals	.512
Organized for learning goals	.433
Links to prestigious sponsor	.297
Overall	.622

Research Question 2: *How do teacher IA project ratings compare with* OPERA*?*

Table 5 shows, for each indicator, the % of IA projects where *all* three teachers gave a '5' rating, the % of IA projects identified as exhibiting the indicator by the 'out-of-the-box' OPERA and by the trained OPERA, and the overall correlation between the median of the teachers' ratings and trained OPERA's probability score. The indicators are ranked by teacher ICC values (see Table 4). For example, for the "content seems appropriate for age range" indicator, teachers showed high agreement between their ratings (.717); teachers' ratings indentified 81.5% of the IA projects as possessing that indicator; the 'out-of-the-box' OPERA identified 95% of IA projects as possessing that indicator; and trained OPERA indentified 85% as exhibiting that indicator; finally, the correlation between median teacher ratings and the probability score is moderate at .56.

TABLE 5.
Teacher and Opera classifications, and correlation between the two.

Quality Indicator	**Has indicator: teachers**	**Has indicator: out-of-box OPERA**	**Has indicator: trained OPERA**	***r* between median teacher rating and trained OPERA**
Content seems appropriate for age range	81.5%	95%	85%	0.56
Identifies age range	94%	3%	79%	0.49
Has instructions	91.5%	84%	85%	0.44
Identifies learning goals	79.5%	3%	44%	0.44
Organized for learning goals	84%	95%	95%	0.72
Links to prestigious sponsor	53.5%	0%	5%	0.12

From Table 5, we note that with training, OPERA's identification of indicators increases, and better matches those of the teachers. Trained OPERA shows good similarities in classification on two indicators: "content seems appropriate for age range"

and “has instructions”. Correlations between human and machine ratings are also good for those two indicators. Both versions of OPERA appear to show little discrimination power on two indicators: “organized for learning goals” and “links to prestigious sponsors”. Recall that these two indicators also showed the lowest ICC values between the teachers, and it may be that OPERA has no strong model upon which to base its decision.

A multiple regression was run to determine the degree to which the OPERA’s confidence scores reflected teacher ratings (1-5), and the influence of the six different categories. The independent variable was the median teacher rating, while dependent variables included OPERA confidence score (0-1), and the rating category. A total of 660 ratings (110 for each category) were used in the model. The final regression ($R^2 = .33$, $F(3, 657) = 108.09$, $p < .01$) suggests three variables are predictive of teacher ratings at a statistically significant level. OPERA score ($\beta = 2.15$, $p < .01$) is positively related to teacher ratings, and prediction improves when ratings are for either “content seems appropriate for age range” ($\beta = 0.92$, $p < .01$) or “has instructions” ($\beta = 0.27$, $p < 0.02$). In terms of practical significance, about 33% of the variability in teacher ratings is accounted for by these two ratings categories, and the Opera score. Given the variability in teacher ratings themselves, this is a substantial amount of predictive power.

Research Question 3: *To what extent can* OPERA *be used to select ‘quality’ IA projects?*

As noted, trained OPERA’s output better matched the teachers than ‘out-of-the-box’ OPERA, especially on two indictors (“content seems appropriate for age range” and “has instructions”). These were also indicators with good teacher agreement. Thus, trained OPERA could perhaps be used to automatically detect quality along these two indicators.

OPERA’s output was poor on “links to prestigious sponsor”, but this indicator also had the lowest teacher ICC value, suggesting a lack of agreement among teacher raters. Thus, is an indicator with little agreement among teachers a reasonable measure of quality? We believe it may simply be too subjective, and should not be left to an algorithm to determine.

SIGNIFICANCE

This paper has addressed the thorny and complex problem of measuring quality in peer produced products. Following previous research, we choose to not define quality as a binary “thumbs-up/thumbs-down” construct; rather we acknowledge quality is defined by a confluence of sometimes subjective indicators. We also investigated whether a machine learning algorithm, OPERA, can serve as a proxy for the laborious and tedious task of assessing quality in online resources for the purpose of supporting and facilitating teacher peer production.

Like previous research, we found that human raters sometimes agree and disagree about quality once decomposed into different indicators. Indicators with little agreement among teachers should not be used as measures of quality. For at least two quality

indicators with lots of teacher agreement, trained OPERA's performance showed moderate levels of agreement with teacher judgments. With future enhancement, it may be possible to use trained OPERA as part of the peer production process.

REFERENCES

Bethard, S., Wetzler, P., Butcher, K., Martin, J. H., Sumner, T. (2009). Automatically Characterizing Resource Quality for Educational Digital Libraries. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries.* New York: ACM.

Benkler, Y. (2006). *The Wealth of Networks: How Social Production Transforms Markets and Freedom:* Yale University Press. New Haven, CT. 528 pp.

Computing Research Association. (2005). *Cyber-infrastructure for Education and Learning for the Future: A vision and research agenda*. Washington, D.C.

Fitzgerald, M.A., Lovin, V., & Branch, R.M. (2003). A Gateway to Educational Materials: An Evaluation of an Online Resource for Teachers and an Exploration of User Behaviors. *Journal of Technology and Teacher Education*, *11*(1), 21-51.

Lagoze, C., Van de Sompel, H., Nelson, M., Warner, S. (2002). Open Archives Initiative - Protocol for Metadata Harvesting - v.2.0. Retrieved February 26, 2010 from http://www.openarchives.org/OAI/openarchivesprotocol.html

Lamantia, J. (2003). Analyzing Card Sort Results with a Spreadsheet Template. Retrieved January 18, 2008, from Boxes and Arrows Web site: http://www.boxesandarrows.com/view/analyzing_card_sort_results_with_a_spreadsheet_template

Maurer, D. & Warfel, T. (2004). Card Sorting: a definitive guide. Retrieved January 18, 2008, from Boxes and Arrows Web site: http://www.boxesandarrows.com/view/card_sorting_a_definitive_guide

McArthur, D. J., & Zia, L. L. (2008). *From nsdl 1.0 to nsdl 2.0: towards a comprehensive cyberinfrastructure for teaching and learning*. Paper presented at the Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries, Pittsburgh PA, PA, USA.

McMartin, F. (2004). MERLOT: A Model for User Involvement in Digital Library Design and Implementation. Journal of Digital Information, 5(3).

Pea, R., with Borgman, C. L. C., Ableson, H., Dirks, L., Johnson, R., Koedinger, K., et al. (2008). Fostering learning in the networked world: The cyberlearning opportunity and challenge, a 21st century agenda for the National Science Foundation. *NSF Report, June*.

Recker, M., Dorward, J., Dawson, D., Halioris, S., Liu, Y., Mao. X., Palmer, B., Park., J. (2005). You Can Lead a Horse to Water: Teacher development and use of digital library resources. In *Proceedings of the Joint Conference on Digital Libraries,* ACM, New York, NY

Smith, M., & Casserly, C. (2006). The Promise of Open Educational Resources. Change: The Magazine of Higher Learning, 38(5), (Sept/Oct), pp. 14.

Sumner, T., Khoo, M., Recker, M., & Marlino, M. (2003). Understanding educator perceptions of "quality" in digital libraries. In *Proceedings of the Joint Conference on Digital Libraries*, NY: ACM.

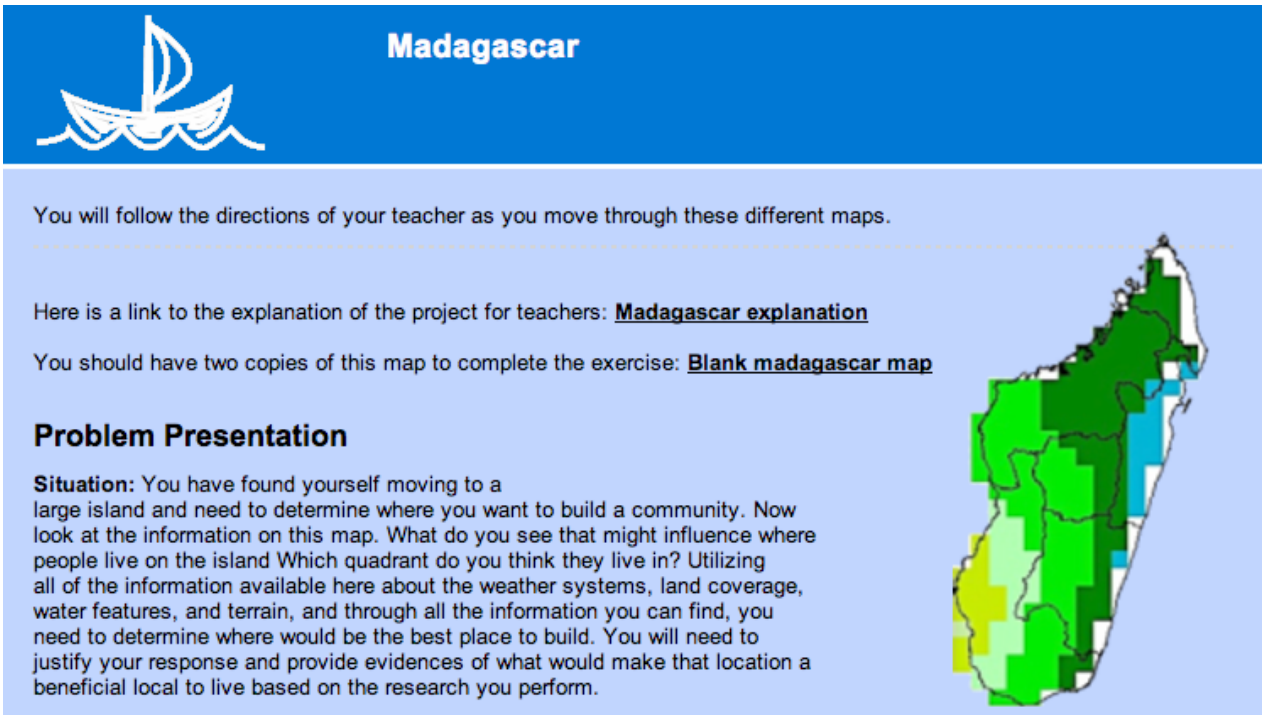

FIGURE 1
Example IA project, showing teacher annotations (text) and embedded online resource.