Do Faculty Agree on Learning Object Selection?

Russell Walker
rwalker2@devry.edu
College of Business and Management
DeVry University
United States

Abstract: Effective reuse of shared educational resources from learning object repositories depends on the ability of faculty to select relevant learning objects for inclusion in specific courses. There has been little investigation to date into faculty decision-making processes for learning object selection. This preliminary analysis explored the degree of consensus among faculty course developers as to which learning objects are relevant to a course. In three pairs of faculty course developers, each member of a pair independently judged the relevance of learning objects to a common course. Agreement between members was low for all three pairs (Cohen's $\kappa = 0.20$, 0.39, and 0.09 respectively), suggesting that substantial differences may exist in how individual faculty members choose learning objects. A larger study using the same methodology is planned to more definitively address the question of faculty agreement on learning object selection.

Introduction

Learning objects are modular digital resources such as images, audio and video clips, tutorials, and simulations that can be reused for teaching and learning (Wiley, 2002). For some time, there has been considerable interest in reusing such shared resources to reduce the cost and enhance the quality of courses in higher education (Metros & Bennett, 2002; Sammons & Ruth, 2007). Recently, there has been a specific focus on open educational resources (OERs), materials that can be freely reused and repurposed because they are in the public domain or were published with permissive intellectual property licenses (Allen & Seaman, 2012). (The categories of OERs and learning objects overlap substantially but are not synonymous, as some learning objects are not freely licensed and thus do not qualify as OERs.)

An apparent disconnect continues to exist between the perceived value of these shared resources and their actual use in courses. For example, Allen and Seaman (2012) found that among academic leaders in higher education, most are aware of OERs, 57% believe OERs have value for their institutions, and two-thirds agree that OERs can reduce costs; yet those same academic leaders reported limited use OERs, with only one-half reporting that even a single course at their institution incorporates any OERs. Allen and Seaman's survey determined that administrators and individual faculty members are the predominant "gatekeepers" for OER adoption, with individual faculty playing the leading role in most institutional categories. Faculty members, in turn, identified the difficulty of finding and evaluating OERs as the primary barrier to wider use of these resources. These findings by Allen and Seaman suggest that investigating how faculty find, evaluate, and choose digital resources may clarify the issues surrounding OER/learning object adoption. This paper describes a preliminary analysis, using data from a study of learning object recommender systems, that explores one key aspect of the resource selection process: the extent to which different faculty members agree when choosing learning objects for a course.

Reuse Models and Selection

Several models of the learning object lifecycle have been proposed. In all these models, the selection of the learning object for reuse in a specific instructional context plays a critical role. Collis and Strijker (2004) proposed a basic six-step learning object lifecycle consisting of (1) obtaining, (2) labeling, (3) offering, (4), selecting, (5) using, and (6) retaining. Collis and Strijker defined the selecting step as one of deciding which of many available learning objects are usable for the application at hand, and stated that such factors as the advice of colleagues, advertising,

vendor contracts, and trade show exhibits could influence the selection process. Collis and Strijker emphasized the complexity of the selection step, including the need to consider content, tone, style, presentation, granularity, cost, ownership, and other criteria.

Sampson and Zervas (2011) sought to integrate several learning object lifecycle models and extend them with defined roles for participants and a comprehensive workflow. Sampson and Zervas specified participant roles for teachers, authors, instructional designers, learning object repository managers, and learners. Additionally, Sampson and Zervas refined Collis and Strijker's (2004) six-step model into a more detailed fourteen-step sequence consisting of (1) identify educational needs, (2) search, (3), develop, (4) describe, (5) offer, (6) approve, (7) publish, (8) select, (9) obtain, (10) modify, (11) integrate, (12) use, (13) feedback, and (14) delete. In the Sampson and Zervas model, the select step is performed by those in the role of either teacher or instructional designer, and the selection should be based on criteria established in the earlier identify instructional needs step. Like Collis and Strijker, Sampson and Zervas listed several factors that could influence selection, including comments from those in other roles, evaluations such as peer reviews, cost, and intellectual property restrictions.

Prior Research on Learning Object Selection

Surprisingly little research was found in the literature relating to the decision process used by faculty members to select educational resources for a course. In particular, little appears to be known about the key question of whether individual faculty members reach similar judgments about whether a given learning object is relevant to, and could therefore be used in, a given course. The question of agreement on selection of learning objects is important for several reasons. The potential time- and cost-saving benefits of learning objects are based on the premise that a resource developed by one educator for one course can be reused by other educators in similar courses. If determinations of what is relevant to a course differ substantially among educators, the likelihood that a learning object developed by one individual will be selected for adoption by many other individuals in similar circumstances may be greatly reduced. Additionally, if learning object selection decisions are highly individual, this may limit the value of such factors as peer reviews, expert evaluations, and usage statistics as guides in selecting learning objects.

A study by Cechinel and Sanchez-Alonso (2011) comparing expert reviews and user ratings of learning objects in the MERLOT repository may be used to draw some inferences about the consistency of learning object evaluations. MERLOT, one of the largest and best-known repositories, provides two distinct types of evaluation for its materials: peer reviews by designated experts, and user ratings which may be contributed by any community member. Cechinel and Sanchez-Alonso compared these for the subset of MERLOT materials with both types of evaluations, and found only low levels of agreement. Cechinel and Sanchez-Alonso offered two possible explanations for this result: either experts and community members use different criteria in evaluating learning objects, or they use similar critieria but rate objects differently against these criteria. Notably, Cechinal and Sanchez-Alonso only compared evaluations between the expert and community member groups; they did not examine consistency of evaluations within each group. Thus, it is impossible to tell whether the inconsistencies they observed between experts and community members might also exist between individuals within these groups.

Method

The analysis described in this paper took advantage of data available from an evaluation of learning object recommender systems (Walker 2012) to conduct a preliminary investigation of the extent to which faculty members agree on selection of learning objects for a course. This was done to determine whether a larger-scale investigation of the topic might be warranted, and to pilot data analysis procedures for such an investigation. In the recommender system study, learning object recommendations were generated for 46 faculty course developers using several recommendation algorithms. The faculty course developers evaluated whether each recommended object was relevant to their most recently developed course. The primary objective of the study was to compare the effectiveness of the recommendation algorithms. However, it happened that in three cases, recommendations were generated for two faculty developers who had worked on the same course at different times. In each of these cases, many recommended learning objects were the same for both developers. This provided an opportunity to examine, albeit with a limited data set, whether the faculty members in each pair reached similar relevance judgments on these learning objects for the common course.

All faculty members participating in the study had developed one or more courses in the business program for a single large U.S.-based for-profit university system within the previous three years. The courses for which two different developers participated were an undergraduate course in Business Intelligence and graduate courses in Business Law and Sustainability Marketing. In each case, the two course developers had worked independently on the common course at different times within the preceding 3-year period. The course developers in each pair were based at different, widely separated campus locations. For the Business Intelligence course, both developers were full-time faculty; for the Business Law course, one was part-time and one was full-time; and for the Sustainability Marketing course, both were part-time.

Learning objects evaluated by the faculty developers were drawn from a data set of 3,392 English-language materials in the Business category of the MERLOT repository having an intended audience of College or Graduate School students. Some recommendations were based on matching learning object descriptions to interest profiles provided by developers and to keywords extracted from course syllabuses. Recommendations were also generated by matching interest profiles of developers and MERLOT community members and recommending objects rated highly by those MERLOT members. Finally, the 10 highest-rated learning objects in the MERLOT business category were recommended to all developers as a control. For each of the three pairs of developers with common courses, there were 23-26 objects that were recommended to both developers in the pair. (The algorithm used to generate each recommendation was not considered in the analysis described here, as the focus here was on how the developers evaluated the common learning objects once they were presented.)

The study design incorporated features expected to enhance consistency of participants' relevance judgments. Participants were subject matter experts in the fields in which they were asked to make assessments, a factor shown to improve agreement among relevance judges in other contexts (Bailey et al. 2008). Participants were given a standard definition of relevance adapted from one developed for the well-known Text Retrieval Conference (TREC) and widely used in information retrieval research (U.S. Department of Commerce 2006). Relevance assessment was done in the context of a simulated work task as recommended by Borlund (2003).

Analysis of the data followed the procedures described by Hartmann (1977) for evaluation of trial-based interobserver reliability. Simple percentage agreement of relevance assessments was calculated for each pair, as this is the most commonly reported statistic for summarizing agreement of binary (yes or no) observations. This is the percentage of cases in which both developers agreed that the learning object was either relevant or nonrelevant. Additionally, because simple percentage agreement may be viewed as distorted by agreement on the large number of nonrelevant objects in each set, effective percentage agreement for relevant objects only was calculated. This is the percentage of cases in which both developers agreed an object was relevant, out of all cases in which at least one developer judged the object relevant. Finally, Cohen's kappa statistic (κ) was calculated; this can be seen as a percentage agreement corrected for the probability that agreement may have occurred by chance.

Results

Results for the three pairs of developers are presented in 2x2 matrix form in Tables 1, 2, and 3. Interobserver reliability statistics for all three pairs are summarized in Table 4.

		Developer B	
		Nonrelevant	Relevant
Developer A	Relevant	1	2
	Nonrelevant	17	7

Table 1: Numbers of relevant and nonrelevant learning objects as judged by each developer for Business Intelligence course.

		Developer B	
		Nonrelevant	Relevant
Developer A	Relevant	6	3
	Nonrelevant	14	0

Table 2: Numbers of relevant and nonrelevant learning objects as judged by each developer for Business Law course.

		Developer B	
		Nonrelevant	Relevant
Developer A	Relevant	3	2
	Nonrelevant	15	6

Table 3: Numbers of relevant and nonrelevant learning objects as judged by each developer for Sustainability Marketing course.

Course	Number of Learning Objects (N)	% Agreement	Effective % Agreement (Relevant Objects)	Cohen's κ
Business Intelligence	27	70%	20%	0.20
Business Law	23	74%	33%	0.38
Sustainability Marketing	26	65%	18%	0.09

Table 4: Summary of interobserver reliability statistics for three courses.

Discussion

Although the simple percentage agreement values in each case appear to indicate at least moderate consistency between the course developers, this is misleading, as it primarily reflects agreement on the large number of nonrelevant objects in each set. In contrast, the effective percentage agreement for relevant objects and the Cohen's κ both indicate very low consistency in how the course developers in each pair identify relevant objects. A Cohen's κ of greater than 0.60 is generally regarded as good interobserver agreement among behavioral researchers (Hartman 1977). The κ values seen here fall far short of that standard. Thus, at least in this small data set, there is evidence of substantial disagreement between two course developers as to which learning objects are relevant to the same course.

This limited analysis provides little basis for proposing an explanation for these disagreements. In general, it appears that in each case, one developer was substantially more liberal in his or her relevance assessments, judging 1.5 to 3 times as many learning objects as relevant than did the other developer in the pair. As was the case in Cechinal and Sanchez-Alonso's (2011) study, it is unknown whether this reflects different relevance criteria used by the faculty developers, different application of similar criteria, or both.

Certainly, the data set in this small sample, obtained as a by-product of a larger study, is insufficient to draw any firm conclusions. However, in combination with Cechinal and Sanchez-Alonso's (2011) findings of little agreement between expert and community member ratings in MERLOT, it does suggest a need to further investigate the level of agreement or disagreement between faculty members as to learning object relevance, as a potential window into the critical process of how educators select resources for courses.

Next Steps

A larger-scale study is currently planned in which multiple faculty members will be asked to select learning objects for common courses from sets of identical recommended objects. Relevance assessments will be analyzed using the procedures described in this paper to measure interobserver reliability. Also, more detailed quantitative and qualitative data will be collected on faculty members' evaluation criteria to gain additional insight into the selection process.

In the planned study, a sample of courses will be randomly selected from the catalog of a large multi-campus university system. For each course, a list of recommended learning objects will be generated based on key phrases extracted from the course syllabus by a machine learning algorithm, as described by Walker (2012). Faculty members teaching different sections of the course on different campuses will be asked to independently assess whether each recommended learning object is relevant (i.e, suitable for use in the course). Faculty members will also be asked to rate each learning object on a scale of 1 (worst) to 5 (best) in three categories used for peer reviews in

MERLOT—quality of content, teaching effectiveness, and ease of use (Cechinel & Sanchez-Alonso 2011)—and will be asked to briefly explain in their own words why they marked each object as relevant or nonrelevant.

Interobserver reliability as measured by percentage agreement, effective percentage agreement, and Cohen's κ will be calculated for the overall relevance assessments and for each of the three category ratings. Faculty members' open-ended explanations of their relevance assessments will be examined using qualitative content analysis methods to determine underlying sources of agreement or disagreement. Findings from these analyses may shed additional light on the consistency of faculty decision-making when selecting learning objects for a course.

References

Allen, I. E., & Seamann, J. (2012). *Growing the curriculum: Open education resources in U.S. higher education.* Quahog Research Group, LLC and Babson Survey Research. Retrieved from http://www.onlinelearningsurvey.com/reports/growingthecurriculum.pdf

Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A. P., & Yilmaz, E. (2008). Relevance assessment: Are judges exchangeable and does it matter? In S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, & M.-K. Leong (Eds.), *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 59-66). New York: Association for Computing Machinery. doi:10.1145/1390334.1390447

Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research, 8(3).* Retrieved from http://informationr.net/ir/

Cechinal, C., & Sanchez-Alonso, S. (2011). Analyzing associations between the different ratings dimensions of the MERLOT repository. *Interdisciplinary Journal of E-Learning and Learning Objects, 7*.

Collis, B., & Strijker, A. (2004). Technology and human issues in reusing learning objects. *Journal of Interactive Media in Education, 2004(4),* 1-32. Retrieved from http://jime.open.ac.uk/

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis, 10*(1), 103-116.

Metros, S. E., & Bennett, K. (2002). Learning objects in higher education. *EDUCAUSE Center for Applied Research Bulletin, 2002*(19). Retrieved from http://www.educause.edu/ecar/

Sammons, M. C., & Ruth, S. (2007). The invisible professor and the future of virtual faculty. *International Journal of Instructional Technology and Distance Learning, 4(1).* Retrieved from http://www.itdl.org/

Sampson, D. G., & Zervas, P. (2011). A workflow for learning objects lifecycle and reuse: Towards evaluating cost effective reuse. *Educational Technology and Society, 14(4),* 64-76. Retrieved from http://ifets.info/journals/14_4/7.pdf

U.S. Department of Commerce, National Institute of Standards and Technology. (2006). *Data – English relevance judgements.* Retrieved from Text Retrieval Conference website: http://trec.nist.gov/data/reljudge_eng.html

Walker, R. (2012). *Comparing information retrieval effectiveness of learning object recommendation strategies for course developers* (Doctoral dissertation). Northcentral University, Prescott Valley, AZ.

Wiley, D. A. (2002). Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. In D. A. Wiley (Ed.), *The instructional use of learning objects* (pp. 3-23). Bloomington, IN: Agency for Instructional Technology and Association for Educational Communication and Technology.