Paradata as Potential Indicators of OER Utility

Marcia A. Mardis
Florida State University
USA
mmardis@fsu.edu

Chandrahasa R. Ambavarapu
Florida State University
USA

Abstract: Open educational resources (OER) are heralded in a global movement toward high quality, affordable, accessible, and personalized education. However, stakeholders have expressed concern about scaling OER use due to a lack of means to ensure fit between learner, resource, and task. A possible way determine fit is through examination of transactional data, i.e., paradata, such as reviews, ratings, views, downloads, and "favoriting." We examined National Science Digital Library (NSDL) paradata, the largest extant accessible corpus, for the extent to which resource fit can be determined from user- and system- generated data. We conducted sentiment analyses of user reviews and correlations between the sentiment scores and data elements. While some relationships between paradata elements were discernible notions of resource fit, reliable research in this area depends on access to larger and more robust paradata sets. We conclude with observed data trends and further research directions.

Introduction and Overview

Efforts to promote common learning standards and shift to digital textbooks to make K-12 education more affordable and accessible are rapidly increasing open educational resource (OER) demand (Ash, 2012; Mickey & Meaney, 2013). However, policymakers, educators, and learners are concerned about an adequate supply of readily available, quality learning resources in freely accessible repositories, registries, and other online communities of practice (Byrne, 2003; Lagoze, 2010; McMartin et al., 2008), particularly in the areas of science, technology, mathematics, and engineering (STEM) learning (Porcello & Hsi, 2013). To determine resource quality and appropriateness, educators may need to rely on more than resource descriptions (i.e., metadata) alone to judge suitability. While attaching metadata to learning resources has been a primary way to support OER selection and implementation and descriptive metadata may help to determine what a resource is about, educators also benefit from information about how resource has been used and whether others have found the resource to be effective (U.S. Department of Education, 2012, 2013).

Metadata are data about the data; paradata are usage interaction data, that is, the data around the data (MacNeill, Campbell, & Hawksey, 2014). Practitioners of learning analytics have the potential to use these generated by the educational resource repository users to recognize notable patterns in digital resource use and impact. Because paradata include system generated as well as social contributions such as ratings, reviews, comments, favorites, downloads, and other activity data relating to resource use, their analysis can give insight into users' perceptions and preferences (Campbell & Barker, 2013; U.S. Department of Education, 2012).

A popular way of deriving meaning from free text user-provided feedback is through opinion mining via sentiment analysis. Sentiment analysis has grown out the need for providers like shopping site Amazon.com (Liu, Yu, An, & Huang, 2013), travel planning site Trip Advisor (Lak & Turetken, 2014), and restaurant review site Yelp (Ganu, Kakodkar, & Marian, 2013; Seaman & Allen, 2014) to be able to be able to recommend products and services to their users. This emerging area of text mining and computational linguistics may also provide a useful, but largely unexplored, approach to examining learning resource paradata. Complementary machine learning work has been done to assess quality from metadata based on expert reviews (e.g., Bethard, Wetzler, Butcher, Martin, & Sumner, 2009); however, very little work has been done to determine ways in which user contributed and user generated paradata might reflect perceptions of resource quality and use. In this study, we first analyzed learning resource paradata with sentiment analysis to explore how useful affective feedback is for developing and curating collections of quality assured web-based resources useful for teaching and learning. We then attempted to determine the possible relationships between sentiment and other use data.

Research Questions

RQ1. To what extent are users' assessments of resource quality positive, negative, or neutral?
RQ2. What are the relationships between user sentiments and other learning resource paradata?

Literature Review

Teaching and learning with digital resources can engage, challenge, and motivate students with compelling experiences that encourage a culture of independent and collaborative learning. Engagement and motivation are achieved through a complex mix of aesthetic, technical, and educational learning resource design. When learners have the ability to use and create digital resources to extend their repertoires, they are able to realize learning outcomes; support their agency; develop metacognitive skills and higher order thinking skills; and participate in reflection and collaboration. While effective teaching is a key element in fostering learner outcomes, in the immediate future, so is the availability of high quality, engaging, personalized learning content (Hanover Research Council, 2011; Maull, Saldivar, & Sumner, 2010; New Media Consortium [NMC], 2014).

Learning personalization results from data-informed decisions made at key points in the instructional process: determination of teacher and student characteristics; guidance of a coherent, rigorous curriculum; execution of appropriate student-centered teaching activities; and application of meaningful formative assessments (U.S. Department of Education, 2012). This process rests on a foundation of high quality, engaging openly accessible learning resources, or OER (Collins & Levy, 2013; Hewlett Foundation).

Open Educational Resources

Due to decreasing state expenditures for instructional materials and growing state mandates for digital textbooks (Collins & Levy, 2013; Mickey & Meaney, 2010, 2011; Simba Information, 2014) and widespread adoption of common learning standards, OER are a growing part of educators' classrooms (Digital Textbook Collaborative, 2012; Mardis & Everhart, 2013; Mickey & Meaney, 2013) and school library (American Association of School Librarians [AASL], 2011; Project Tomorrow, 2012; School Library Journal, 2013) learning resource base. The Hewlett Foundation defines OER as "teaching, learning, and research resources that reside in the public domain or have been released under an intellectual property license that permits use and repurposing (para. 3)."

OER originated in and are predominantly created for K-12 STEM learning (Hanson & Carlson, 2005; Mardis, 2003; Mardis & Howe, 2010; Mardis & Zia, 2003). While the benefits of using OER versus costly commercial instructional materials are evident, problems of determining OER quality have persisted for over two decades (Hewlett Foundation, 2013; Okerson, 2000). In an unpublished report of a 2013 meeting of statewide education officials hosted by the National Science Digital Library, participants stated a dependence on and a fear of losing an adequate supply of high quality STEM OER. Participants cited threats such as a lack of curation strategies to manage collections; insufficient resources to identify and vet OER; and a strong desire to automate the selection, validation, and management processes as essential issues to resolve if their delivery of common standards curricula were to be successful (National Science Digital Library [NSDL], 2013). Many researchers have noted that metadata standards are neither necessary nor sufficient to address issues to expressing learning resource quality to users (Bethard et al., 2009; Porcello & Hsi, 2013; Wetzler et al., 2013) because OER quality often includes issues of curation that are not captured in descriptive metadata schema.

The curatorial enterprise is comprised of resource selection, management, and promotion (Rosenbaum, 2011). In recent years there has been a growing awareness that paradata, such as ratings and user reviews, are needed to properly curate the content in repositories and determine resource suitability for teaching and learning tasks, as Griffin (2013) pointed out, "Subjective reviews submitted from highly qualified educators as well as independent reviewers are valuable elements that could be, and should be, included as unique data..." (para. 3), and features that capture other aspects of the user experience such incidences as favoriting, sharing, viewing, and downloading (Campbell & Barker, 2013).

Traditional instructional materials are primarily reviewed for content accuracy, particularly in the STEM fields (Spiegel, 1989; Stern & Roseman, 2004). As materials have become more digital and curriculum standards more influential, these media quality and standards alignment issues must also be routinely considered (Hanson & Carlson, 2005; Mardis, ElBasri, Norton, & Newsum, 2012). However, in an autonomous search mode, when

teachers assess quality "on the fly," sentiment and quality determination tends to vary widely (Perrault, 2007; Recker et al., 2011; Recker et al., 2007), with some teachers favoring assessments based on appearance and some favoring assessments based on content (Price, 2007). For this reason, many teachers and education policymakers have expressed preferences for repositories of vetted resources with clear designations of quality, utility, and curriculum support (Griffin, 2013; Sumner, Khoo, Recker, & Marlino, 2003).

However, encouraging teachers to contribute plentiful, complete, and accurate reviews of resources has proven to be difficult, despite the fact that many teachers report a favorable response to complete reviews left by other educators. Incentives are rarely effective because many teachers view feedback about their resource preferences as personal criticism. Teachers tend to be altruistically motivated and enthusiastic about sharing information about a resource when they feel the resource is valuable and would help other teachers (Van Acker, van Buuren, Kreijns, & Vermeulen, 2013).

Method

In this study, we analyzed paradata from resources in the National Science Digital Library (NSDL), a leading provider of K-12 OER. NSDL began providing paradata for resources in its collections in 2011 in the form of annotations about the resources and related information on usage (VanGundy, 2010). NSDL's two distinct paradata schema (comm_anno and comm_para) (National Science Digital Library [NSDL], 2014a, 2014b, 2014c), are considered among preferred formats by many OER providers (Bienkowski, Brecht, & Klo, 2012; Campbell & Barker, 2013; Niemann, Scheffel, & Wolpers, 2012; U.S. Department of Education, 2014).

Description of the Data

NSDL maintains all paradata records in the form of two types of XML record comment sets: "comm_anno" XML sets contain review text and "comm_para" XML sets contain non-textual paradata such as star ratings, favorites, and downloads. Table 1 lists the collections that provide paradata to NSDL.

The researchers harvested the XML paradata files directly from NSDL servers and them combined the comm_anno and comm_para files to integrate all available paradata for each resource. In sum, we extracted a total of 2505 comm_anno and comm_para resource records from NSDL. Table 2 provides an overview of the paradata included in the integrated file. As Table 2 shows, the sample included 752 comments. Records also included 1232 other occurrences of user-provided (subjective) in the form of favorites, features, ratings, and recommendeds as well as 521 system-generated (objective) data in the form of downloads and views, for a total of 2505 discrete paradata.

We used the Simple API for XML (SAX) Java API to parse the XML documents. The parser extracted data for paradata specific fields (such as user comments, views, and downloads in our case) and excluded content from other fields. The parsed paradata were redirected to an Excel spreadsheet using the Poor Obfuscation Implementation (POI) API.

Procedure

We performed our analysis in two phases to establish sentiment polarity (positive or negative) and its relationship to subjective rating (e.g., "two and a half stars") and system-generated paradata.

Phase I. Sentiment Analysis. We used sentiment analysis to determine comments whether uses reviews and comments were positive, negative, and neutral and assigned a numerical weight according to each determination. As the base for our sentiment analysis, we used commonly accepted techniques (Padmaja & Fatima, 2013) including the Natural Language Toolkit (NLTK), an external library for Python, because it can perform a vast amount of text processing and analysis. We used a NLTK-trained text processing classifier derived from the data set created by Pang and Lee (2005). This classifier contains data from movie and Twitter reviews in which the reviews are categorized as positive, negative and neutral. For resources that had multiple reviews, we calculated an average (i.e., arithmetic mean) of sentiment classification numbers. Once the sentiment analysis scores were assigned, both researchers independently reviewed random selections of the polarity assignments and achieved approximately 90% agreement with the machine assignment.

Phase II: Bivariate Correlation. Using Pearson Product Moment Correlation analysis, the annotation sentiment average values (N=501) were first correlated with subjective paradata (favorites, features, ratings, and

recommendations), then with objective paradata (downloads, views). The researchers chose to use Pearson correlational approach because it is well suited to exploring relationships. Finally, the subjective and objective paradata were correlated absent the average sentiment values. All correlation analyses were conducted using the Statistical Package of the Social Science (SPSS).

Validity

The limitation with sentiment analysis is that machine learning for this kind of problem set is in its preliminary phase and yet to undergo further research to achieve 100% accuracy. Subtle language and domain-specific language can affect sentiment accuracy. Neutral reviews required internal parsing to determine polarity. There are certain user comments like "5 star" or "cool" and some comments that begin with strong positive sentiment and ends with weak negative sentiment, for which emotion calculated is neutral, although they are considered positive. After consulting research by experts in sentiment analysis on how to address this issue (Wilson, Wiebe, & Hoffmann, 2009), we determined that manual review was needed for the analysis of such comments to determine true expressed sentiment and reclassify them accordingly. Finally, it should be noted that the prior research done in this area suggests that automated sentiment analysis works best with big datasets and cannot be relied upon for smaller data samples (Pang & Lee, 2008). Despite these potential challenges, each researcher checked different random samples of the polarity assignments and found the sentiment analysis to be 90% accurate.

Results

The NSDL file reflected the textual and non-textual paradata collected for 2505 records. Of the records in that file, 757 had user-contributed annotations, i.e., comments or reviews. These reviews contained 13353 words of resource review text, with an average of 18 words per annotation and approximately two sentences per annotation. Figure 1 depicts the distribution of comments across records.

Phase 1. Sentiment Analysis Results

The first step of the sentiment analysis was to examine all of the annotation entries and determine their polarity. Then, we divided the range of sentiment scores into three equal ranges: negative, neutral, and positive. Figure 2 provides an overview of the distribution of the annotations across polarity.

As Figure 3 depicts, the data set included 135 neutral annotations with sentiment scores ranging from -.147 to .460, with a mean of .042 and a median of .000. The 49 negative annotations had sentiment scores ranging from -.754 to -.146, with a mean of -.364 and a median of -.404. Most of the annotations were positive (n=579). Positive sentiment scores ranged from .461 to 1.07, with a mean of .508 and a median of .500. For resources that had more than one review (n=501), we then calculated the arithmetic means of the sentiments for those resources to enable further explorations. Figure 3 depicts the distribution of average annotation sentiment means relative to a normal curve. As Figure 3 shows, the distribution of the average annotation sentiment somewhat fits the normal curve, with a skewness of -.956 and a kurtosis of 1.064. Negative values for the skewness indicate data that are skewed left, with the left tail is long relative to the right tail. A positive kurtosis indicates a "peaked" distribution, with few extremely high values.

Next, we plotted the individual sentiment scores versus the average sentiment scores, as Figure 4 depicts. When the average annotation sentiments were plotted against the individual sentiment scores, the results showed that, while the averaging muted very high and very low values, the majority of the average scores remained clustered similar to the pattern of the individual scores, thus suggesting that they would be viable for use in correlational analyses to detect linear relationships.

Phase 2. Bivariate Correlations

Next, we calculated Pearson Product Moment Correlation coefficients among the 6 non-textual system-generated paradata and average sentiment scores. Using the Bonferroni approach to control for Type I error across the bivariate correlations, a *p* value of less that .01 was required for significance. Table 3 depicts the results. The results of the analyses presented in Table 3 show that four correlations were statistically significant. The most significant correlation was between Featured (i.e., the number of times the resource was a featured resource on the website) and Downloaded at p=1.00. Other significant correlations were between Viewed and Favorited (p=.384) as

well as Viewed and Featured (p=305.) and the Average Sentiment and Rating (p=.158). Although not statistically significant, negative correlations were observed between Viewed and Downloaded (p=-.372), Featured and Favorited (p=-.285), and Rating and Favorited (p=-.059). Average Sentiment and Rating demonstrated very few notable relationships with objective or subjective paradata elements.

Discussion

The findings, when compared to the literature foundation, provided insight on the research questions:

RQ1. To what extent are users' assessments of resource quality positive, negative, or neutral?

The literature relating to user reviews suggests that although users most frequently leave comments that express extremely positive or extremely negative sentiment, comment readers and users value comments that include both positive and negative assessment elements. Our word clouds revealed that NSDL commenters tended to use a balanced approach because the sentiments primarily contained resource description words and secondarily contained resource use words. Most reviews appeared to be about the format of the resource, i.e., "lessons" and "videos." The fewest annotation words related to the content of resources such as "water" and "Hemisphere."

Our paradata analyses suggested that while users tended to leave positive comments more frequently, those positive comments tended to be weakly positive, weakly negative, or neutral. These kinds of assessments suggest that NSDL collection users, presumably educators, blend positive and negative elements in their annotations, along the lines of the "feedback sandwich" form of teacher feedback in which a negative comment in placed between positive comments (Milan, Parish, & Reichgott, 2006; Parkes, Abercrombie, & McCarty, 2013) or the modulated language that consumers tend to employ when reviewing a product they've experienced (Kronrod & Danziger, 2013; Mudambi & Schuff, 2010). Unlike to content of consumer reviews, very few of the annotations reflected a very high sentiment score, which few readers find helpful anyway (Ludwig et al., 2013).

RQ2. What is the relationship between sentiment assessments and resource use?

Bivariate correlational analysis revealed very weak or no correlations between annotation sentiment and indicators of resource use, such as incidences of downloading, viewing, or favoriting. However, the relationship between Featured and Downloaded was significant and strong, suggesting that when a collection provider promotes a resource, that additional attention results in resource use. The power of promotion perhaps accounts for the significant relationship between Featured and Viewed as well.

That Viewed and Favorited had a significant correlation may be symptomatic of a Facebook-like behavior in which a user reviews items quickly and likes (or, here, favorites) any items that seem to appeal (Gerlitz & Helmond, 2013; Sharifrazi & McCabe, 2014); it is difficult to determine the extent to which this relationship proxies an assessment of the resource's quality or usefulness, but that this relationship is more significant than the relationship between Views and Average Sentiment may be telling of a new type of user feedback preference. However, from a curation perspective, it may be important for collection providers to consider targeted promotion of their collections to ensure broad use.

Conclusion

Driven by the common standards movement and federal education accountability requirements, education trends toward personalized learning are prompting an explosion in interest, and even anxiety, about securing a large supply of high quality open educational resources for K-12 learners (Hanover Research Council, 2012; Hewlett Foundation, 2013). In this study, we took an initial look at the extent to which subjective and objective learning resource paradata from the National Science Digital Library, a leading OER provider, can be used alone to determine resource usage, utility, and quality.

Overall, the sentiment analysis revealed only initial insight into users' resource quality perceptions. Reviews tended to be short with brief, declarative sentences that reflected a narrow range of sentiment. Bivariate correlations suggested that more could be learned from the relationship between non-textual paradata then from the relationship between sentiment text and non-textual paradata elements. Whether the usefulness of social media sentiment to make predictions is underappreciated (Gayo-Avello, 2012) or whether NSDL's educational paradata

corpus is currently too small to produce reliable characterizations (Lak & Turetken, 2014), when these conclusions are examined through the lens of General Sentiment theory, they suggest that sentiment analysis of educational paradata is not yet useful to detect valence or affect and sentiment analysis alone may never be sufficient. Possibly, though, Gerlitz and Helmond (2013) may have correctly pointed to natural evolution of user engagement on the web with the growing use of quick, easy, measurable button-generated interactions like ratings and "like." Despite this uncertainty, the researchers sense that this type of research will yield much more interesting and helpful results as the paradata corpus grows in size and quality. Nonetheless, this study did produce some recommendations and directions for further research.

Limitations, Recommendations, and Directions for Further Research

We acknowledge several limitations to this study should be acknowledged and can be addressed with recommendations for content providers and addressed in directions for further research.

1. Beware of "junk" annotations and misspellings. The data set contained annotations such as "That was crazy Jeffrey wrote this" or "b" or "I really love this sight." Content providers can address this issue by providing a review scaffold with character limits and spell check features. Researchers should be sure to check for misspellings and inappropriate reviews and flag the annotations for further review.

2. Be attuned to subtle language and overly declarative annotations. Because research has suggested that users to tend to soften their language when writing reviews, sentiment analysis may underrepresent subtle reactions such as "so educational" and "thank you for making this video." Likewise, sentiment analysis may overweigh hyperbolic comments such as "this is the coolest thing I've ever seen." These issues are especially problematic in the analysis of a mixed sentiment comment such as "Very simple. I used as a homework assignment. The students wanted more variety and more challenge, though." Content providers can scaffold with annotation prompts instead of free text boxes that will guide users' contributions with specific questions. Scaffolds can also prompt for comments relating to content and teaching strategies. Content providers can also employ annotation-voting features to allow users to mark reviews they find helpful.

3. Researchers can improve and expand the dictionaries they use to detect sentiment and use word, phrase, and concept level analyses (Cambria, Schuller, Liu, Wang, & Havasi, 2013). Researchers can also consider a discipline-specific dictionary that reflects aspects of resource quality drawn from instructional materials selection and teaching literature.

4. Factor in the user's identity. User profiles were not part of this study and user job role, experience level, and work site may have a relationship with review content. Collection providers may wish to consider gathering minimal user profile information to gain a sense of the contributor. Researchers may wish to explore differences between collections that contain discipline expert annotations and collections with teacher annotations as well as compare expert and teacher annotation ratings, downloads, and other non-textual paradata. Expert input shows great promise as a basis for machine learning techniques to automatically assess resource quality (Bethard et al., 2009; Wetzler et al., 2013); the extent to which sentiment analysis can factor into improving or extending automatic assessment is a fertile area for exploration.

5. Factor in the context of contribution. Many reviews appear to have been left during the course of a professional development event or content provider presentation. To the extent possible, collection providers may wish to capture the context in which the review is contributed, perhaps with a checkbox or radio button selection in the annotation authoring dialogue. Researchers may wish to analyze reviews for indicators of application such as "Students examined this video in small groups…"

The power of paradata to inform learning personalization may not be ready to be realized, but as the field of learning analytics continues to grow as a result of, or along with, the rising use of OER and their social media features, this type of large scale educational data analysis offers teachers, students, and researchers an unprecedented opportunity to use an understanding of resource use to personalize learning.

Acknowledgement: This work was supported in part by IMLS RE-04-09-0055, NSF #1304382, NSF #1449622.

Tables and figures can be viewed at http://bit.ly/2eNMhI3 ; References can be viewed at http://bit.ly/2f343bn