Automatically Assessing Resource Quality for Educational Digital Libraries

Philipp Wetzler
University of Colorado
594 UCB
Boulder, CO, USA
philipp.wetzler@colorado.edu

Steven Bethard
University of Colorado
594 UCB
Boulder, CO, USA
steven.bethard@colorado.edu

Kirsten Butcher
University of Utah
1705 Campus Center Dr
Salt Lake City, UT, USA
kirsten.butcher@utah.edu

James H. Martin
University of Colorado
430 UCB
Boulder, CO, USA
james.martin@colorado.edu

Tamara Sumner
University of Colorado
594 UCB
Boulder, CO, USA
tamara.sumner@colorado.edu

ABSTRACT

With the rise of community-generated web content, the need for automatic assessment of resource quality has grown. We demonstrate how developing a concrete characterization of quality for web-based resources can make machine learning approaches to automating quality assessment in the realm of educational digital libraries tractable. Using data from several previous studies of quality, we gathered a set of key *dimensions* and *indicators* of quality that were commonly identified by educators. We then performed a mixed-method study of digital library quality experts, showing that our characterization of quality captured the subjective processes used by the experts when assessing resource quality. Using key indicators of quality selected from a statistical analysis of our expert study data, we developed a set of annotation guidelines and annotated a corpus of 1000 digital resources for the presence or absence of the key quality indicators. Agreement among annotators was high, and initial machine learning models trained from this corpus were able to identify some indicators of quality with as much as an 18% improvement over the baseline.

Categories and Subject Descriptors

H.3.6 [**Information Systems**]: Library Automation; H.3.7 [**Information Systems**]: Digital Libraries—*Standards, User issues*; I.2.7 [**Computing Methodologies**]: Natural Language Processing—*Text analysis*; I.5.4 [**Computing Methodologies**]: Applications—*Text processing*

WICOW'09, April 20, 2009, Madrid, Spain.

General Terms

Human Factors, Algorithms

Keywords

digital library, digital resource, machine learning, natural language processing, quality

1. INTRODUCTION

In recent years, "quality" has emerged as a critical, yet poorly understood concern on the World-Wide Web. This concern has grown in importance with the rise of user- and community-generated content such as Wikipedia articles, digital library resources, product reviews, answer forums, etc. As such user-generated contributions grow in importance, the need to automatically ascertain the quality of these contributions increases. To date, a wide variety of approaches for automatic quality analysis have been suggested, from examining word counts [3] to link structure [23] to changes in revision histories [1, 27].

The work described here addresses quality in the context of educational digital library efforts such as the National Science Digital Library (NSDL) and the Digital Library for Earth System Education (DLESE). One goal of these libraries is to develop and curate collections of web-based resources useful for teaching and learning across a wide range of grade levels and educational settings, including primary, secondary and tertiary education, and in both formal and informal learning settings. These resources include textual materials (background readings, references), interactive and visual materials (maps, animations, simulations), classroom and laboratory activities, and scientific data. They are created by individuals and institutions across a range of academic, government and non-profit sectors, and submitted to these libraries for further dissemination into educational settings. As such, vetting the quality of these resources is a critical issue for educational libraries, but one that is very challenging in practice to carry out reliably and efficiently at the necessary scale. It is not unusual for libraries to receive requests to curate collections containing 1000s of web-based teaching and learning materials.

Within these efforts, library developers engaged in resource selection and collection curation processes are increasingly being tasked with designing and managing collections to reflect specific library policies and goals aimed at promoting quality. Concerns about the quality of library resources often revolve around issues of accuracy of content, appropriateness to intended audience, effective design and information presentation, and completeness of associated documentation or metadata. Such quality evaluations require making difficult, complex and time-consuming human judgments to assess whether resources belong in particular collections or libraries. These judgments are influenced by a variety of factors, for example, the information present in the resource, structural and presentational aspects of the resource, and knowledge about the resource creators. Thus, there is a critical need in educational digital libraries for cognitive tools [21, 25] to support library developers, and ultimately library users, to more effectively and efficiently assess the quality of educational resources.

The goal of the work presented here is to address this need in a scalable way by training computational models that can automatically recognize the characteristics of high quality digital resources. These models are not meant to produce simple yes-no decisions, but to give detailed information to a user about where the resource is good and where it needs work. For example, a resource might have a clear introduction and identify important learning goals, but at the same time use terminology too difficult for the 6th grade audience it claims to target. Simply assigning this resource a classification of “high” or “low” quality would be unhelpful, and perhaps even misleading. With more detailed feedback however, a digital library developer could make an informed choice about whether to include the resource in a collection, and a learner could better decide whether the resource fits their educational needs. In general, the goal of this project is to develop computational models that can identify the important characteristics of resource quality and use these to help scaffold human judgments. These algorithms can underpin not just tools for curators and library developers, but for end-users such as teachers and learners as well.

2. RELATED WORK

Prior work suggests that quality is a complex and multidimensional construct. Many qualitative and quantitative studies of human subjects have explored how users make judgments about the quality of online information sources. For example, Fogg and colleagues conducted an online survey of over 1400 participants to identify what affects perceptions of web site credibility [16]. They found that factors like ease-of-use, expertise and trustworthiness were most important to their participants. In another study, Rieh used verbal protocols and post-search interviews to study how people make judgments related to quality and cognitive authority while searching for information on the web [22]. Rieh found a slightly different set of factors that were important for this task, including accuracy, currency, trustworthiness, scholarliness and authoritativeness. Focus groups conducted by Sumner and colleagues identified yet another set of criteria for quality judgments when engaged in collection curation within educational digital libraries [24]. These included scientific accuracy, lack of bias, and good pedagogical design. While a number of common themes run through the dimensions of quality identified by all these studies, the variations among the studies reflect the difficulty of characterizing quality in a concrete, usable way.

Several recent studies were able to show that certain low-level characteristics of digital resources correlated well with overall judgments of quality. For example, Ivory et al. showed that low-level design issues, such as the amount and positioning of text, or the overall portion of a page devoted to graphics, correlated highly with expert judgments of overall site quality [17]. Using a slightly different approach to the problem, Custard and Sumner trained machine learning models to judge overall quality using low-level features like website domain names, the number of links on a page, how recently a page was updated, and whether or not videos or sound clips were present [7]. To get quality judgments for their resources, they had digital library personnel judge digital resource collections as either high, medium or low quality. Their models were able to identify whether a resource was from a high quality, medium quality or low quality collection with 76.67% accuracy. These results are encouraging as they suggest that overall quality may be characterized using a selection of low-level features. Work still remains to identify a characterization of quality that is more easily adapted to the various different perspectives on quality that are required by a variety of users and usage scenarios.

Algorithms for assessing quality in its many definitions have been broadly proposed ever since web usage entered mainstream society. In addition to simple keyword matching and exploiting the link structure of the web (e.g. PageRank [4]) to identify important and relevant documents, quality metrics such as the time since last update of a web page, presence of broken links, and the amount of textual content have been shown to improve results on internet search [28]. More recent research has attempted to ground web page quality in existing information quality models, and generally found that quality of web pages needs to be considered in the context of the domain or application on one side and the targeted audience on the other side [2], [19].

The popularity of the online encyclopedia Wikipedia has prompted research into its overall quality of information, and more specifically into evaluating the quality of individual articles. The Wikipedia community maintains a set of *featured articles*, which are considered to be the best that Wikipedia has to offer; they are selected by community review according to a set of well defined criteria (e.g. *well written*, *factually accurate*, *follows Wikipedia style guidelines*) [26]. Stvilia et. al. used textual features like *readability* and metadata features such as *number of editors* and *age of the article* in combination with machine learning techniques to predict *featured article* status of Wikipedia articles with relative success [23]. Blumenstock later showed that the task can be accomplished with extremely high accuracy using only the word count of an article as a feature [3].

Metrics based on simple text statistics such as number of syllables in words and average sentence length, which are easily computed automatically, have a long history of being used in an educational context to assess a text’s reading level, and more recently for automatic essay grading. Concrete formulations of these ideas are the Flesch Reading Ease Score and the Flesch-Kincaid Grade Level. More recent efforts attempt to automatically assess readability and textual coherence and cohesion using computational semantics [13].

All of these efforts suggest a desire to characterize quality using more concrete features and metrics. A key component of our research is to identify the characteristics of quality that humans rely on most.

3. UNDERSTANDING QUALITY

Prior research which tried to define the term "quality" found widely differing components of quality depending on the particular scenario in which they studied it. Thus, to make *quality* more tractable to a computational approach, it was crucial to operationalize the definition in a way that was more concrete, designed with a specific application in mind — in this work we focused on educational digital libraries — and that was agreed upon by human users.

We relied on two major sources of information. First, we performed an analysis of data from a number of prior studies on the quality of educational digital resources, gathering a list of common dimensions and indicators of quality. Second, we interviewed experts in digital library management, presenting them with resources and the dimensions of quality, and asking them to perform quality related tasks. The results of these two studies provided the kind of information necessary to formulate a concrete, learnable definition of quality.

The following sections describe the analysis of prior studies and our expert study in more detail.

3.1 Analysis of Prior Studies

Prior research efforts have investigated how educators characterize the quality of digital resources [16, 17, 24], revealing a number of common dimensions of quality, such as scientific accuracy, lack of bias, and good design or organization. To get a clearer picture of how educators characterize quality, an analysis was performed using the raw data collected by three studies:

Educator Reviews for DWEL The Digital Water Education Library [10] encourages review of their resources by their educational community. We gathered all the educator comments provided in the full reviews for resources targeted at grades 9-12, for a total of 364 reviews generated by 21 reviewers for 182 unique URLs.

Educator Reviews for Climate Change Collection The Climate Change Collection [6] was developed using an interdisciplinary review board for selecting appropriate high quality resources. We obtained all the narrative comments concerning digital resource quality from 55 individual reviews provided by 4 individuals for the 28 grade 9-12 digital resources in the collection.

Educator Focus Groups In 2002, Sumner and colleagues hosted a series of focus groups where science educators discussed the quality of digital library resources [24]. We acquired the transcribed verbal data generated by 38 educators as they reviewed 18 resources.

The qualitative verbal data collected from all three of these studies was then coded by two raters, first to filter out comments that were not relevant to quality, and then to derive the most important dimensions of quality indicated by the data. The latter was performed in an iterative process where comments were grouped by similarity into categories, and then the categories were iteratively adjusted until they best covered the data. Priority was given to categories that were identified in all three sources of data, and categories were adjusted until 100% interrater agreement was reached.

The result of this analysis was a set of 25 dimensions that described the approaches to quality taken by educators across all the different studies. Based on the frequency with which each dimension was observed in the studies, the 12 top dimensions listed in Table 1 were selected.

Because of the bottom-up iterative approach taken to identify these dimensions, the study also resulted in a list of low-level *features* or *indicators* for each dimension that identified some of the more concrete and observable factors that encompassed the more conceptual dimensions of quality. For example, the "Appropriate pedagogical guidance" dimension had indicators like "Has instructions" and "Identifies learning goals", while the "Age appropriateness" dimension had indicators like "Identifies age range" and "Content is appropriate for age range".

Overall, the 12 dimensions above accounted for over 80% of all the comments about resource quality in all of the studies. This level of coverage was clearly encouraging, but since these dimensions were derived from analyzing the comments of educators, it remained to be shown that they were in fact generally applicable for characterizing quality.

3.2 Expert Study

To confirm the importance of our dimensions of quality, we performed a mixed-method study of digital library experts. We recruited a team of eight experts in collections development who were experienced in assessing the quality and appropriateness of digital resources for various digital collections. We presented these experts with digital resources and asked them both to talk about how they would make a quality judgment, and to assign some numeric values to their quality assessments.

First, each expert thought aloud [15] while evaluating the quality of six digital learning resources. The resources were taken from the Digital Library for Earth System Education (DLESE) [9], a repository of digital educational resources about Earth science. They were selected to include both resources that had been peer-reviewed and identified as being high quality, and resources that had been rejected from DLESE for being of too low quality. While the experts examined their six resources, their comments about the positive and negative aspects of the resource were recorded. Then they were asked to give the resource an overall rating, from -3 or very low quality, to +3 or very high quality. Finally, they were asked to make an accept/reject decision, that is, they were asked to decide whether the resource was high enough quality to be included in a digital library collection.

Using each expert's ratings of the six resources, three resources were selected for a more detailed review. These resources corresponded to the expert's highest rated resource, the expert's lowest rated resource and the resource closest to the expert's mean rating. The expert then evaluated these resources using the 12 dimensions of quality identified in the analysis of prior studies, and again their positive and negative comments were recorded, as well as their ratings of how well each dimension was addressed (from -3 to +3).

The products of this study were several hours of verbal data including comments about positive and negative aspects of resource quality, as well as the numerical assessments for overall quality judgments, quality dimension judgments, and inclusion or exclusion from a digital repository.

To assess whether or not the dimensions of quality identified from the previous studies of educators were also used

Dimension	Overall	Accept
Provides access to relevant data	0.67**	0.46*
Good general set-up	0.65**	0.73**
Appropriate inclusion of graphics	0.61**	0.53*
Robust pedagogical support	0.59**	0.48*
Appropriate pedagogical guidance	0.57**	0.52*
Reflects source authority	0.56**	0.54*
Readability of text	0.54**	0.40
Appropriateness of activities	0.49*	0.53*
Focuses on key content	0.42*	0.32
Age appropriateness	0.41*	0.26
Inclusion of hands-on activities	0.36	0.43*
Connections to real-world applications	0.36	0.27

Table 1: The 12 targeted dimensions of quality and their relationship to digital library experts' assessments of overall quality and accept/reject decisions (*$p < .05$, **$p < .01$). The dimensions are ordered by their correlations with overall quality.

by digital library collections experts, we performed several analyses of these data. We looked at how the dimension level quality ratings compared to both the overall quality ratings and to the accept/reject decisions. Table 1 shows these comparisons. The "Overall" column indicates the correlation between the expert's dimension ratings and their overall resource quality ratings. The "Accept" column indicates the correlation between the dimension ratings and the decision to accept or reject the resource from a digital library. 10 of the 12 dimensions were significantly correlated with the overall quality judgments, and 8 of the 12 dimensions were significantly correlated with the accept/reject decisions. There were some differences between which dimensions were most useful for overall quality judgments and which were most useful for accept/reject decisions, but for example, both "Good general set-up" and "Appropriate inclusion of graphics" were near the tops of both lists.

These high correlations demonstrate that our identified dimensions of quality are sufficient to capture the subjective processes that digital library experts use when assessing resource quality. This is encouraging because it means that the dimensions of quality we identified as being used by educators are also used by digital library experts when analyzing the content of digital resources on the web.

3.3 Indicators of Quality

The expert study confirmed that quality could be decomposed into meaningful dimensions that were more concrete than the abstract concept of "quality". However even our list of the most important 12 dimensions included fairly abstract dimensions like "Good general set-up". To make a computational approach to quality feasible, it was necessary to push the decomposition of quality further, identifying low-level *indicators* of quality that were concrete, easily recognizable, and known to be useful to experts of digital resource quality.

Fortunately, candidate indicators for each dimension of quality were already identified in the analysis of previous studies. Thus, the main remaining goal was to identify which such indicators were most important in defining quality. To answer this question, we analyzed the verbal data collected from the digital library experts. All of the spoken data recorded during the assessments of overall quality were hand-coded to identify any time where an expert mentioned a quality indicator as being either present or absent in the resource. Counts for the indicators identified by each of the digital library experts were then tabulated.

Indicator	Correlation
Has prestigious sponsor	0.905
Content is appropriate for age range	0.889
Has sponsor	0.858
Identifies learning goals	0.842
Has instructions	0.755
Identifies age range	0.728
Organized for learning goals	0.612

Table 2: The seven most predictive indicators and their correlations to accept/reject decisions.

Using this data, we were able to look at which indicators were most predictive of the decision to accept or reject a resource from a digital library. We extracted the indicators where both the presence was highly correlated with acceptance and the absence was highly correlated with rejection. Table 2 shows the top seven such indicators. Not surprisingly, one of the most reliable indicators of resource quality was the presence of a prestigious sponsor, such as NOAA, NASA or USGS. Other important cues included tailoring the resource content to a specific age range, and giving guidance on using the resource through instructions and identified learning goals.

These indicators provide a concrete definition of quality which corresponds strongly to the processes experts use in assessing quality. They provide a set of characteristics that identify the conceptual pieces of a resource that are likely to be considered when judging the quality of a resource. In addition, they provide a means of characterizing quality in terms of low-level features that should be more amenable to computational approaches. The following sections examine this last claim in detail.

4. COMPUTATIONAL APPROACH

One of the main goals of our research is to computationally assess the presence or absence of the quality indicators in a given resource with an accuracy that approaches human performance. Similar to prior efforts to determine semantic properties of text, we employ supervised machine learning algorithms to build a statistical model of the available data. Using such a model, judgments can be made about previously unseen resources. This approach requires a training set of digital libraries, some of which exhibit each indicator and some of which don't. We built such a corpus using human annotation.

Our test bed is the DLESE Community Collection (DCC). The DCC is a collection of interdisciplinary resources within DLESE with a general focus on "bringing the Earth system into the classroom". Criteria for inclusion focus on pedagogical value and accessibility in addition to scientific correctness. Another defining characteristic of the collection within DLESE is that it includes resources that were submitted for review by individual DLESE users[11]. All submitted resources are currently manually reviewed by committee for compliance with the standards of DCC content. After the decision has been made to include a resource in the collection, it is then annotated with various metadata describing the new catalog entry.

In the following sections we describe the protocol and results of the annotation project and the machine learning setup we used to create computational models of the quality indicators, as well as present preliminary results.

4.1 The Annotation Project

We selected 1000 earth system educational resources directed at high school students; 950 were selected randomly from DCC, and 50 were selected from those resources rejected by DLESE. When a reviewer decides to reject a resource, they write a short free-form note explaining their reasons. Common reasons for rejection besides quality-related problems are: the resource is outside the scope of DLESE; the type of the resource is one not cataloged by DLESE; or the resource suggested is already in the catalog. Based on the reviewer's notes we only selected resources that were rejected for what appeared to be quality-related reasons.

Two people with previous experience annotating DLESE resources for metadata were asked to judge the presence or absence of the seven quality indicators on each resource. In order to achieve reliable results we carefully formulated instructions for annotation, outlining our definitions for each indicator using concrete terms and examples taken from our expert study. After a short test-run and in cooperation with DLESE experts we made some minor revisions to these annotation guidelines.

Each annotator was then asked to independently look at 600 of the 1000 resources; they were presented with the home page of the resource and allowed to navigate freely. Every resource was annotated at least once, and 200 were double-annotated to allow us to measure agreement between annotators. Low agreement on an indicator either indicates that the annotation guidelines are too inexact, thus letting each annotator come up with their own interpretation, or that annotating that indicator is inherently difficult for people. The natural language processing research community commonly assumes that higher agreement between annotators means a machine learning system will likely be able to achieve better performance. Table 3 shows inter-annotator agreement for each indicator, as well as the percentage of resources where the indicator was marked as present. Agreement was above 80% for 6 of the 7 indicators, suggesting that our guidelines were clear and our characterization of quality was not too subjective.

We also recorded the URLs of all web pages that the annotators visited during their review and that they considered to be a part of the resource. DLESE only stores the URL of first entry into a resource; but many resources consist of multiple linked pages. Identifying the extents of a resource is a complex problem in itself [14], [12], and one that we're not addressing in this work.

After determining the inter-annotator agreement we asked both annotators to look again at the resources and indicators where they disagreed, and discuss and resolve their disagreement. The resulting set of 200 resources can be assumed to be of a higher quality of annotation. These resources will serve as the *test set* and will be used only for the final evaluation of the quality indicator models.

The corpus contains 950 resources that were randomly selected from DCC, and 50 resources that were not allowed into DLESE for quality reasons. Thus, the 950 DCC resources should be of higher quality than the 50 rejected resources. Using our training corpus we evaluated how predictive the indicators are of the "accepted into DCC" status. We also did a leave-one-out evaluation, showing the relative contribution of each indicator. The results of this study can be seen in Table Table 4. These experiments were run on a reduced training set, selected to have an equal number of high quality and low quality resources.

Quality indicator	present in	agreement
Has instructions	39%	85.2%
Has sponsor	97%	99.5%
Has prestigious sponsor	34%	63.6%
Identifies age range	20%	87.3%
Not inappropriate for age	99%	100.0%
Identifies learning goals	28%	83.1%
Organized for goals	76%	80.6%

Table 3: Quality indicator presence in resources and inter annotator agreement

	accuracy
all indicators	71%
w/o *Has instructions*	−5%
w/o *Has sponsor*	−2%
w/o *Has prestigious sponsor*	−16%
w/o *Identifies age range*	−7%
w/o *Not inappropriate for age*	−2%
w/o *Identifies goals*	−4%
w/o *Organized for goals*	−4%

Table 4: Quality indicator predictiveness and leave-one-out analysis

The seven quality indicators together were able to accurately predict whether a resource was ultimately accepted into DCC with an accuracy of 71%. This is encouraging, as it shows that the quality indicators truly capture relevant aspects of quality. It also leaves room for improvement; an automatic system for assessing quality for a specific task may want to introduce additional indicators to improve performance.

The other take-away from Table 4 is the relative importance of the quality indicators. Some, such as *has sponsor*, contribute little (because almost *all* resources had a sponsor, so it doesn't serve well as a distinguishing feature, see also Table 3). The indicator that contributed most by far is *has prestigious sponsor*; it was also the most difficult and subjective, as the agreement numbers show.

4.2 The Computational Models

The quality indicators are assessed at the level of an entire resource, e.g. an entire resource is considered to either have instructions or not, and an entire resource is considered to be age-inappropriate or not. Thus every classification decision we make looks at a complete resource – containing multiple web pages and possibly rich media and linked PDF files – as a unit.

In order to classify a resource using the machine learning algorithm we must encode it into a numerical vector. The encoding process should attempt to catch salient features present in a resource that may help in determining the presence or absence of an indicator while discarding information that is too complex for the statistical algorithms. The way in which those features are presented greatly influences how effectively they can be used by the algorithm. Our efforts to find an effective set of features and an effective encoding are guided by a large corpus of previous work in using machine learning on linguistic and semantic tasks; even so the set of features that allow effective models to be built can only be identified experimentally.

Our research platform is based on the ClearTK toolkit for statistical natural language processing [20].

4.2.1 *Feature Extraction*

To build the vectorial representation of a resource that is required by the machine learning system, we extract a number of numerical and *yes/no* features of a document. The following is the feature set used in the system we are reporting on here:

Bag-of-words, bag-of-bigrams This feature set is a common starting point for many natural language processing applications. It simply indicates to the machine learning system if any given word shows up somewhere in the current resource or not. E.g. "resource contains the word 'a'", "resource contains the word 'seismic'", "resource contains the word 'record'", and so on, for every distinct word that a resource contains.[1] *Bag-of-bigrams* does the same for every two consecutive words that occur in the resource.

TF-IDF *term frequency – inverse document frequency*, a refinement of the *bag-of-words* feature that gives words a different weight, based on how often they show up in the current resource vs. all resources. For example, the word "and" will show up many times in all resources, so the feature "resource contains the word 'and'" will be indicated to the machine learning system as not very important. On the other hand, the feature "resource contains the word 'Rayleigh'", assuming the word "Rayleigh" shows up a number of times in the current resource, but almost never anywhere else, will be marked as particularly important.[1]

Resource URL This feature presents the resource URL to the machine learning system. In addition to the full URL we include the domain and super-domains, e.g. "`http://web.ics.purdue.edu/~braile/edumod/surfwav/surfwav.htm`", "`web.ics.purdue.edu`", "`ics.purdue.edu`", "`purdue.edu`", "`edu`". This helps the machine learning system make useful generalizations about the domain a resource is hosted in.

URLs linked to We include all the URLs that a resource links to, presented in the same way.

Google PageRank For all URLs we include a feature that indicates the Google PageRank of the respective site. This indicates the relative importance of that site on the internet, measured by how many other sites link to it. For example, a site like `http://www.nasa.gov/` has a high PageRank value, while, e.g. a largely unknown and small university web site will have a low value.

Alexa TrafficRank Alexa[2] is a company offering traffic statistics on web sites based on analyzing user behavior. For all URLs we include their reported *TrafficRank* in our feature set, which indicates the amount of user traffic a web site receives relative to other sites.

Quality indicator	baseline performance	ML performance
Has instructions	61%	78%
Has sponsor	96%	96%
Has prestigious sponsor	70%	81%
Indicates age range	79%	87%
Not inappropriate for age	99%	99%
Identifies learning goals	72%	81%
Organized for goals	75%	83%

Table 5: Preliminary Results

4.2.2 *The Machine Learning System*

We use the SVMlight package [18], which uses the support vector machine approach to machine learning. This approach has been effective in a wide range of natural language processing applications, using features similar to the ones used here. The training parameters are chosen using cross validation. The results reported below were achieved using a linear kernel SVM.

4.2.3 *Preliminary Results*

We trained and evaluated models on the training data using cross-validation, then compared the results to a simple majority-class baseline. For example, the *has instructions* indicator is present in 39% of resources. If we always assumed that a resource has no instructions, we'd be correct in 61% of cases. An effective machine learning model will show significant improvement over this baseline. Table 5 shows the results of this evaluation.

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. Using the current feature set we were unable to improve performance over the already high baseline on *has sponsor* and *not inappropriate for age*. Our current features do not appear to be sufficient to determine if a resource identifies its learning goals.

These results are encouraging in that even using very basic features we are able to classify some of the indicators fairly well, and they are guiding our current efforts to grow the feature set by adding features that specifically address some aspects we are currently missing.

5. DISCUSSION & FUTURE WORK

5.1 Expert Study

The encouraging performance of our computational models of quality is in part a validation of our methodology for selecting quality indicators, in which we relied heavily on the study of human processes. Our expert study took the very general ideas of quality suggested by prior research, and explored these ideas with quality assessment experts to produce both qualitative and quantitative data. These data verified that the dimensions of quality derived from the analysis of prior work were in fact dimensions that quality experts used frequently. Perhaps more importantly however, these data also allowed us to see how expert attention to particular indicators of quality compared with their decisions to

[1] These examples are taken from `http://web.ics.purdue.edu/~braile/edumod/surfwav/surfwav.htm`

[2] `http://www.alexa.com/`

accept or reject a resource from a collection. This allowed us to select a set of concrete indicators of quality that were highly correlated with the kind of judgments human experts were making. As the computational results showed, by using these indicators as part of a more concrete definition of quality, we were able to make automated methods for characterizing quality more tractable.

Our work here demonstrates the importance of considering human processes in developing computational models. This approach seems to be particularly helpful for areas like quality where the tasks are sometimes vague or poorly defined. These task definitions can be better refined by collecting data about how humans approach the tasks, and then using statistical analyses to isolate the key components of the human processes. The resulting components not only encourage more concrete task definitions, but also offer benefits for computational approaches.

5.2 Quality Indicators

The annotation project has highlighted the issue of quality indicator distribution within our test environment: Some indicators (most notably *has sponsor* and *not inappropriate for age*) show a very uneven distribution, as almost all of the resources exhibit those indicators. This is a problem for the statistical processes of the machine learning system, as they rely on having many examples of both sides (indicator present and absent) to find reliable ways to distinguish between the two cases. In order to make progress on those indicators an extended data set will probably be necessary.

It will be interesting to see if including other quality indicators which the research community has looked at will help an overall quality assessment. One such indicator could be *text cohesion* or *readability scores* (see for example Coh-Metrix[13]). These do not require any further annotation, as previous work has been done to identify approaches that correlate well with human judgment.

5.3 Computational Models

We have presented results that show the feasibility of modeling quality indicators with natural language processing and machine learning techniques. Current results show success on some of the indicators, including the ones that appear to be more relevant for overall quality according to a preliminary study, but performance is still poor on others. The current feature set offers little access to the semantics or the argument structure of a resource. In order to improve performance across all indicator models we intend to explore a larger set of features that aims to capture the structure of the content and to identify the more important concepts within a resource, rather than treating all parts as equally important. In particular we are pursuing two directions:

5.3.1 Surface Structure

Resources that are cataloged by DLESE and other libraries are for the most part in HTML format, potentially linking to PDF files or containing rich media. Currently we use an HTML parser and a simple ad-hoc rule system to extract the text portions of a page, and discard parts that we don't need, such as scripts. The extracted text is noisy: it still contains many things that are not part of the textual content of the web page. In particular it doesn't attempt to distinguish between navigation elements, boilerplate (e.g. page headers or footers, copyright), advertisements, and educational content. A web page offers many visual cues to help the user identify these parts and navigate the text, but in the flat text format we currently use those cues are lost.

We intend to improve on this by building a domain independent system that splits the content into blocks, then classifies those based on textual and HTML cues, to not only identify the non-content parts of a page, but also to split the content into headings and paragraphs. Prior efforts in this area (see for example CLEANEVAL[5]) don't provide the rich structural annotation that we're aiming for and only focus on identifying the textual content.

Having identified those content classes in a resource allows more targeted features. For example, instructions that help the user approach a resource effectively are likely to be found early on in the resource and may be structurally separated from other parts, e.g. consisting of a separate paragraph. With the added information a model for the *has instructions* indicator may be less likely to be distracted by other sections of the resource that use similar terminology.

5.3.2 Semantic Features

The content features used by our system, such as bag-of-words, rely solely on counting words that show up in the training set. This leads to problems when a resource uses slightly different terminology than previously seen resources to describe, for example, learning goals. The automatic system ignores the new words, because they haven't been used when talking about learning goals before; a human reader, on the other hand, could use rich understanding of the words' meaning to recognize that the new words are talking about the same thing. The *Heat Transfer and El NiÃśo* resource mentioned in the error analysis provides a concrete example: the resource referred to learning goals as "curriculum standards", as opposed to other, more common phrases, such as "education standards" or "learning objectives".

On another note, using the current feature set the algorithms see each resource as essentially a large collection of disjointed words, making it hard to distinguish between occasional usages of words like "instruction", and document sections that discuss instructions in a focused way. Lexical methods were used successfully in [8] to identify overarching key concepts within a set of resources. In our future work, we aim to capture more fine-grained, discourse level concepts by using a richer semantic feature set. Having identified key concepts in a paragraph, taking into account the words' semantics rather than just their surface form, the machine learning algorithms should be able to focus on the actual content of the resource versus picking up individual words out of context.

6. CONCLUSIONS

We have presented a principled approach to defining the quality of web resources and training models to perform automatic quality assessments. Through the analysis of prior work and our own expert study, we identified key *indicators* of quality that are both used by experts in quality assessment and easily recognized by non-experts. We constructed a training corpus of 1000 digital resources annotated with these quality indicators, and trained machine learning models which were able to identify important indicators, like the presence of a prestigious sponsor or age range specifications, with accuracies over 80%. These models can underpin tools ranging from quality assessment engines that can help digital library curators manage large collections to end-user tools that can help students learn to better evaluate the quality of resources they see online.

7. ACKNOWLEDGEMENTS

This material is based in part upon work supported by the National Science Foundation under Grant Numbers CISE/IIS-0741326 and CISE/IIS-0534581. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We also wish to thank our colleagues at Digital Learning Sciences / UCAR for their help with the expert study, the test-bed creation, and the error analysis.

8. REFERENCES

[1] B. T. Adler and L. de Alfaro. A content-driven reputation system for the wikipedia. In *Proceedings of the 16th international conference on World Wide Web*, pages 261–270, Banff, Alberta, Canada, 2007. ACM.

[2] S. ann Knight and J. Burn. Developing a framework for assessing information quality on the world wide web. *Informing Science Journal*, 8:159–172, 2005.

[3] J. E. Blumenstock. Size matters: Word count as a measure of quality on wikipedia. In *Proceedings of the 17th International World Wide Web Conference*, pages 1095–1096, New York, NY, USA, 2008. ACM.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Computer Networks and ISDN Systems*, pages 107–117, 1998.

[5] CLEANEVAL home page. `http://cleaneval.sigwac.org.uk/`, Oct. 2008.

[6] Climate change collection. `http://serc.carleton.edu/climatechange/`, Oct. 2008.

[7] M. Custard and T. Sumner. Using machine learning to support quality judgments. *D-Lib Magazine*, 11(10), Oct. 2005.

[8] S. de la Chica. *Generating Conceptual Knowledge Representations to Support Students Writing Scientific Explanations*. PhD thesis, University of Colorado, 2008.

[9] Digital library for earth system education. `http://www.dlese.org/`, Oct. 2008.

[10] Digital water education library. `http://www.csmate.colostate.edu/DWEL/`, Jan. 2004.

[11] DLESE Community Collection (DCC) scope statement. `http://www.dlese.org/Metadata/collections/scopes/dcc-scope.php`, Oct. 2008.

[12] P. Dmitriev. As we may perceive: Finding the boundaries of compound documents on the web. In *Proceedings of the 17th International World Wide Web Conference*, 2008.

[13] D. F. Dufty, D. Mcnamara, M. Louwerse, Z. Cai, and A. C. Graesser. Automatic evaluation of aspects of document quality. In *Proceedings of the 22nd annual international conference on Documentation*, 2004.

[14] N. Eiron. Untangling compound documents on the web. In *Proceedings of the 14th ACM Conference on Hypertext and Hypermedia*, pages 85–94, 2003.

[15] K. A. Ericsson and H. A. Simon. *Protocol Analysis: Verbal Reports as Data*. The MIT Press, revised edition, Apr. 1993.

[16] B. J. Fogg, J. Marshall, O. Laraki, A. Osipovich, C. Varma, N. Fang, J. Paul, A. Rangnekar, J. Shon, P. Swani, and M. Treinen. What makes web sites credible?: a report on a large quantitative study. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 61–68, Seattle, Washington, United States, 2001. ACM.

[17] M. Y. Ivory, R. R. Sinha, and M. A. Hearst. Empirically validated web page design metrics. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 53–60, Seattle, Washington, United States, 2001. ACM.

[18] T. Joachims. *Making large-scale support vector machine learning practical*, pages 169–184. MIT Press, 1999.

[19] M. J. Kargar, A. R. Ramli, H. Ibrahim, F. Azimzadeh, and S. B. B. M. Noor. Assessing quality of information on the web towards a comprehensive framework. *Iranian Journal of Engineering Sciences*, 1, 2007.

[20] P. V. Ogren, P. G. Wetzler, and S. Bethard. ClearTK: A UIMA toolkit for statistical natural language processing. In *UIMA for NLP workshop at Language Resources and Evaluation Conference (LREC)*, 2008.

[21] T. C. Reeves, J. M. Laffey, and M. Marlino. Using technology as cognitive tools: Research and praxis. In *Australian Society for Computers in Learning and Tertiary Education (ASCILITE)*, 1997.

[22] S. Y. Rieh. Judgment of information quality and cognitive authority in the web. *Journal of the American Society for Information Science and Technology*, 53:145—161, 2002.

[23] B. Stvilia and M. B. Twidale. Assessing information quality of a community-based encyclopedia. In *Proceedings of the International Conference on Information Quality*, pages 442–454, 2005.

[24] T. Sumner, M. Khoo, M. Recker, and M. Marlino. Understanding educator perceptions of ”quality” in digital libraries. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 269–279, Houston, Texas, 2003. IEEE Computer Society.

[25] T. Sumner and M. Marlino. Digital libraries and educational practice: a case for new models. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 170–178, Tuscon, AZ, USA, 2004. ACM.

[26] Wikipedia:featured article criteria. `http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria`, Oct. 2008.

[27] H. Zeng, M. Alhossaini, L. Ding, R. Fikes, and D. Mcguinness. Computing trust from revision history. In *Proceedings of the 2006 International Conference on Privacy, Security and Trust*, Oct. 2006.

[28] X. Zhu and S. Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In *SIGIR ’00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 288–295, New York, NY, USA, 2000. ACM.