Journal of Universal Computer Science, vol. 22, no. 1 (2016), 94-113
submitted: 30/6/15, accepted: 20/12/15, appeared: 1/1/16 © J.UCS

Mining Models for Automated Quality Assessment of Learning Objects

Cristian Cechinel
(Federal University of Pelotas, Pelotas, Brazil
contato@cristiancechinel.pro.br)

Sandro da Silva Camargo
(Federal University of Pampa, Bagé, Brazil
sandro.camargo@unipampa.edu.br)

Miguel-Ángel Sicilia
(University of Alcalá, Alcalá de Henares, Spain
msicilia@uah.es)

Salvador Sánchez-Alonso
(University of Alcalá, Alcalá de Henares, Spain
salvador.sanchez@uah.es)

Abstract: The present paper presents the results of an alternative approach for automatically evaluating quality inside learning object repositories that considers lower-level measures of the resources as possible indicators of quality. It is known that current repositories face a difficult situation, as their amount of resources tends to increase more rapidly than the number of evaluations provided by the community of users and experts. Alternative approaches for automatically assessing quality can relieve human-work and provide temporary quality information before more time and consuming evaluation is performed. We propose a methodology to automatically generate quality information about learning resources inside repositories with Artificial Neural Networks models. For that, we considered 34 low-level measures as possible indicators of quality and we used available evaluative metadata inside two world recognized repositories (MERLOT and Connexions) as baseline information for the establishment of classes of quality. The preliminary findings point out the feasibility of such an approach and can be used as a starting point in the process of automatically generating internal quality information about learning objects inside repositories.

Keywords: Ranking mechanisms, quality assessment, ratings, learning object repositories, artificial neural networks
Categories: L.0, L.1.2, L.3.2

1 Introduction

The process of assessing quality of learning resources inside Learning Object Repositories (LORs) usually involves different forms of evaluative metadata [Vuorikari, Manouselis, & Duval, 2008] provided by their community of users (tags, likes, ratings, comments, lenses). As LORs are proper platforms to foster the creation and strength of ties among people who share interest in the same topics, it is normal that LORs take advantage of such natural structure in order to gather useful quality information about their resources [Han, Kortemeyer, Krämer, & von Prümmer, 2008]. The collected information serves then as an external memory that is used by search and retrieval algorithms to rank and recommend resources to the community of users of the repository, allowing others to find materials in line with their individual needs. Examples of usage of this kind of evaluative metadata can be found in some of the most important LORs existing nowadays, such as: Organic.Edunet (www.organic-edunet.eu), Connexions (cnx.org), OER Commons (www.oercommons.org), and MERLOT (www.merlot.org).

Even though this kind of strategy is efficient to a certain extent, the number of learning resources inside LORs is increasing more rapidly than the user's capacity to evaluate them, thus becoming impossible to depend only on human generated assessments for populate repositories with quality information [Ochoa & Duval, 2009]. The lack of quality information in many resources compromises the process of search and retrieval, as many resources are not able to compete for good rankings positions in comparison with those which were already tagged. Such a problem interferes in the success of the learning object economy that depends on the proper delivering of learning objects (LOs) to complete the whole learning object life-cycle (obtaining, labelling, offering, selecting, using and retaining). The need of providing quality assessment for a wider range of LOs inside LORs calls for the development of tools that can complement existing manual approaches of quality evaluation.

There are already approaches focused on automatically measure some specific aspects of LO quality, however they normally rely on the previous existence of metadata inside LORs, or on information about the resources that can be collected just after the resource is available for a certain amount of time (old resources). Considering that metadata is often inaccurate/incomplete [Cechinel, Sánchez-Alonso, & Sicilia, 2009] [Sicilia, García-Barriocanal, Pages, Martinez, & Gutierrez, 2005] and that it is key to provide quality information also for the recently new arrived resources; we propose an alternative approach for the problem. The intent is to automatically generate quality assessments by using evaluative information inside the repositories together with a set of intrinsic features (i.e. measures that can be directly calculated from the materials) that are possibly associated to quality. The proposed approach was already recently tested in some previous works. For instance, [Cechinel, Sánchez-Alonso, and García-Barriocanal, 2011] developed Linear Discriminant Analysis (LDA) models based on profiles of highly-rated LOs of the MERLOT repository and using 13 intrinsic features of the resources. The resultant models were able to classify resources with accuracies of approximately 72% (for the classification between good versus not-good) and of 91.5% (for the classification between good versus poor). These models were generated for the discipline of *Science & Technology* and the material type *Simulation.* As MERLOT organizes resources in many different categories of discipline (Arts, Business, Humanities, etc.) and material types (Simulation, Tutorial, Animation, etc.), the authors pointed out that models for automated assessment of learning resources should be developed considering the possible combinations among this variety of classifications, and also considering the different groups of evaluators who rated resources (community of users versus community of reviewers). On another experiment, [Cechinel, 2012] evaluated models for three different subsets of MERLOT and encountered models able to classify resources with an overall precision of 89%. However, that experiment was very limited as it worked with a restricted number of resources on each dataset, and also used the entire training datasets to test the models. The present paper extends the previous works in a number of ways. First, here we explore a higher number of subsets of MERLOT repository and use a larger collection of data. Second, while in the previous works the authors explored the creation of highly-rated LOs profiles to then extract the intrinsic features that would be used by the data mining algorithms, here we are using a more algorithmic approach, i.e., the models are being generated exclusively by the use of data mining algorithms.

This article is originally based on a previous paper [Cechinel, Camargo, Ochoa, Sánchez-Alonso, & Sicilia, 2012]. In here, we included a more detailed description of the data used and of the statistical analysis we performed. We also provided a more in depth revision of the literature and most important, we also tested the proposed approach with data collected from a different repository than MERLOT (i.e. Connexions). The rest of this paper is structured as follows. Section 2 describes existing research focused on automated assessment of resources inside repositories as well as existing works presenting intrinsic features of resources that may be associated to quality. The methodology followed for the study is presented in section 3, and section 4 discusses the results found in the experiments conducted with data from MERLOT. Section 5 describes the results of an attempt to extrapolate the proposed methodology to the Connexions repository. At last, conclusions and future work are provided in Sections 6 and 7 respectively.

2 Automated quality assessment of learning objects

[Ochoa and Duval, 2008] proposed a series of metrics to rank LOs during the search and retrieval process inside LORs. The authors used three different aspects of LOs pertinence (personal, topical and situational) together with information gathered from the resources metadata, the user queries, and the records of historical usage of the materials. They have observed a better performance of their approach in comparison to traditional text-based ranking methods. [Koukourikos, Stoitsis, and Karampiperis, 2012] used the comments given by the users of a repository to discover qualitative information about the learning resources. The authors applied sentiment analysis techniques over a set of comments and tested their prediction accuracy using as baseline the ratings available in the repository. Although the experiment is in a very preliminary stage, the authors were able to find results that point to the feasibility of estimating ratings by using user comments as input information. A similar experiment was carried out by [Santos & Cechinel, 2015] who used text mining to predict ratings of LOs based on the comments given by their users.

Interesting works towards automated evaluation can also be found in the domain of digital libraries, precisely related to the evaluation of quality Wikipedia articles. For instance, [Dalip, Gonçalves, Cristo, and Calado, 2011] have recently used several measures of articles (review history, network features and text features) and used them together with machine learning techniques to automatically assess quality inside Wikipedia. The authors have found that the best quality indicators are those that can be directly extracted from the articles, i.e. the textual features. In [Kewen, Qinghua, Yuxiang, & Hua, 2010], the authors built neural networks models and evaluated how 28 different metrics influenced the quality of Wikipedia articles. The authors concluded that lingual features (words count, spell errors, etc.) play an important role in the initial stage of an article quality, and as the quality of an article improves, other features such as the structure (cite numbers, internal links, external links, etc.) and the history (edit times, article age, etc.) become more important.

2.1 Intrinsic features of learning objects

Other works related to Wikipedia can also serve as sources of inspiration for the search of lower level measures that could be associated to LOs quality. For instance, [Blumenstock, 2008] has found associations between the quality of articles in Wikipedia and their respective number of words. He suggested the length of an article can be used as a possible predictor of quality. Moreover, [B. Stvilia, Twidale, Smith, and Gasser, 2005] contrasted highly voted articles in Wikipedia from those not highly voted using the edit history article information and some other article characteristics. According to the authors, the median values of the measures contrasted in the experiment considerably varied between the compared data sets. Examples of those measures are: total number of edits, number of anonymous user edits, number of internal broken links, and number of images. The results of that study and the encountered metrics were latter included on a framework for information quality assessment developed by the authors [Besiki Stvilia, Gasser, Twidale, & Smith, 2007].

Evidences that intrinsic features of resources were associated to quality were also found in the context of educational hypermedia applications. For instance, [Mendes, Hall, and Harrison, 1998] evaluated the sustainability and reusability of such systems by using measures such as, the link type, and the size and structure of the systems. Moreover, [Bethard, Wetzer, Butcher, Martin, and Sumner, 2009] were able to decompose the concept of quality of resources in educational digital libraries into 5 measurable aspects that could be used as indicators of quality and that could be automatically observed. The authors applied natural language processing and machine learning techniques to automatically evaluate the quality of resources. At last, [Ivory and Hearst, 2002] conducted a series of experiments related to the automated assessment of the usability of awarded websites. The authors identified that good websites contain more links and more words than the regular and bad ones.

The approach presented here deals particularly with the features of LOs that are displayed to the users and that are mentioned in the literature as possible associated to quality. Such features are normally related to the categories of *presentation design* and *interaction usability* (used in the Learning Object Review Instrument – LORI [Nesbit, Belfer, & Leacock, 2003]) and the category of *information quality* (usually discussed in the field of educational digital libraries). As will be explained latter, we use the ratings given by the community of experts of MERLOT (peer-reviewers) and the lenses of Connexions as the quality reference for generating our models.

3 Methodology

The goal of the present study is to generate and test data mining models for classifying LOs between *good* and *not-good.* For that, we use only those features that can be calculated directly from the resources themselves, named here as intrinsic features (or low-level measures). In this section we describe the methodology followed in the experiment that uses data from MERLOT repository.

3.1 Data collection

A crawler was used to traverse the pages of MERLOT and collect 35 different metrics of LOs catalogued in the repository. MERLOT was selected due to the large amount of catalogued materials and users, and for implementing a strategy for quality assessment based on evaluations (ratings and comments) given by experts and regular users [Cafolla, 2006], i.e., precisely the kind of information required for conducting our experiments. Considering that LOs catalogued in MERLOT are mainly consisted by websites, we focused on collecting those features that are expected to appear in such kind of material. The features collected are presented in table 1, and are the same as the ones included in the experiments of [Cechinel et al., 2011]. Many of the collected metrics are also included in the works mentioned in the previous section.

Class of Measure	**Metric**
Link Measures	*Number of Links, Number of Unique Links, Number of Internal Links, Number of Unique Internal Links, Number of External Links, Number of Unique External Links*
Text Measures	*Number of Words*, Number of words that are links
Graphic, Interactive and Multimedia Measures	*Number of Images, Total Size of the Images (in bytes), Number of Scripts,* Number of Applets, Number of Audio Files, Number of Video Files, Number of Multimedia Files
Site Architecture Measures	*Size of the Page (in bytes), Number of Files for downloading*, Total Number of Pages

Note. The term Unique means “non-repeated”. The term internal stands for to those links that are located at some directory below the root home-page.

Table 1: Intrinsic features collected for the experiment

Considering that MERLOT materials vary significantly in size, we defined a 2 level depth limit for the crawler. This means to say that metrics were calculated for the home-page of each LOs (root node or level 0), for the nodes linked to the root-node (level 1), and for the nodes linked to the pages on level 1 (level 2). Even though it is possible this limitation can affect the results, this was a required action as the process of collecting information was excessively slow. In total, we collected information from 20,582 LOs. From that amount, only 2,076 were rated by the experts, and 5 of them did not have information about their discipline or material type. The remaining 2,071 were then used in the present study. The frequencies of the LOs for each one of the 105 collected subsets (intersections between category of discipline and material type) are shown in table 2.

Material Type/Discipline	Arts	Business	Education	Humanities	Math & Statistics	Science & Technology	Social Sciences	Total
Animation	4	23	21	16	8	22	4	98
Case Study	0	3	23	16	3	3	2	50
Collection	8	52	56	43	50	80	15	304
Drill and Practice	2	23	13	28	19	37	5	127
Learning Material	5	0	0	0	0	13	0	18
Learning Object Repository	0	0	1	0	4	1	3	9
Lecture/Presentation	6	42	38	48	13	32	20	199
Online Course	0	0	1	0	0	1	0	2
Other Resource	0	0	0	0	0	2	0	2
Professional Paper	0	0	0	0	0	1	0	1
Quiz/Test	0	14	10	4	11	23	1	63
Reference Material	6	83	40	51	68	102	6	356
Simulation	57	63	40	78	40	150	18	446
Tutorial	6	76	73	93	48	86	11	393
Workshop and Training Material	0	0	0	0	0	0	3	3
Total	94	379	316	377	264	553	88	2071

Table 2: Frequency of materials for each subset

As most of the subsets contain a small number of resources, we restricted our study to just a few of them. We decided to work only with 21 subsets that had more than 40 items, and which materials were catalogued in one of the following types: *Collection*, *Reference Material*, *Simulation* and *Tutorial* (grey hashed in table 2). The difficulties for training, validating and testing predictive models for lower subsets (less than 40 items) would be more restrictive. A total of 1,429 LO were included in the experiment, corresponding to 69% of the collected data.

3.2 Classes of quality

Considering that ratings given by peer-reviewers concentrate above rating 3 (see Figure 1), we generate classes of quality the for each subset by using the terciles of the ratings. LO with ratings lower than the first tercile are classified as *poor*, LO with ratings higher than the second tercile are considered as *good,* and LO with ratings in the middle of both terciles are classified as *average.* The terciles for each subset are shown in table 3. The classes of quality *average* and *poor* were then combined to form the *not-good* class.

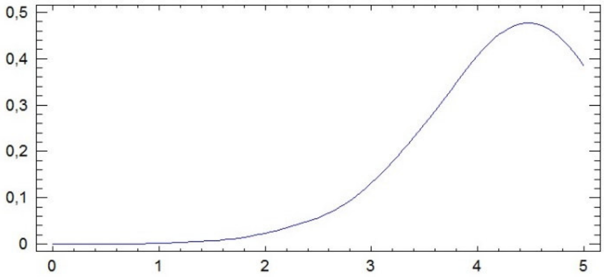


Figure 1: Distribution of ratings among peer-reviewers

Subset	**Arts**	**Business**	**Education**	**Humanities**	**Math & Statistics**	**Science &Tech**
Collection		4.25|4.75	4.25 | 4.75	4.25|4.75	4 | 4.75	4.25 | 5
Reference Material		4 |4.75	4 | 4.75	4 | 4.5	4 | 4.75	4 | 4.75
Simulation	4|4.75	4 |4.75	4 | 5	4 | 4.75	4 | 4.5	4 | 4.75
Tutorial		4 |4.75	4.25 | 5	4.25|4.75	4 | 5	| 4.75

Table 3: Terciles that divide resources into classes of quality for each dataset

3.3 Mining models for automated quality classification of learning objects

Artificial Neural Networks (ANNs) were used to generate and test models for quality assessment of LO using the classes of quality as the output goal and the intrinsic features as the input classes. ANNs are suitable for this kind of problem as they are adaptive, distributed and work well in situations where the pattern among the variables is not known (or not clear). Moreover, ANNs have also shown the best accuracies during a preliminary round of experiments we conducted with different models (rules, decision trees and ANNs). We used the Neural Network toolbox of Matlab with the following settings: 70% of each subset for training, 15% for testing and the remaining 15% for validating, as recommended by [Xu, Hoos, and Leyton-Brown, 2007]. The Marquardt–Levenberg algorithm [Hagan & Menhaj, 1994] was tested using from 1 to 30 neurons. As the subsets were small, and aiming at obtaining more statistically reliable results, we repeated each test 10 times and the average results were computed. The models were created for classifying LO between *good* and *not-good.*

4 Results and discussion

Our models presented distinct performances depending on the subset used for training. Most of models tend to classify *not-good* resources better than *good* ones. This is probably a consequence of the unequal number of resources of each class inside subsets (approximately formed by 2/3 of *not-good* and 1/3 of *good*). These trends can be seen on figure 2 (overall accuracies - lozenges, accuracies for the classification of good resources – squares, and not-good resources - triangles).

In order to observe structural patterns expressed by the models, we ran a Spearman's rank correlation (rs) analysis to test if there were associations between the number of neurons and the accuracies of the models, and to observe the tendencies of these associations for classifying LOs of each class. This is important as it helps one to better understand the complexity behind the classes of quality that are being analysed. For instance, if *x* is a predictive model for a given subset *X*, and *y* is a model for a given subset *Y*; if both have the same accuracies and x has less neurons than *y*, this means to say that patterns existent in *X* are simpler than patterns expressed in *Y*. In other words, it is easier to understand what is good (or not-good) in the subset *X*. Table 4 shows results of such analysis.

Subset	**Arts**	**Business**	**Education**	**Humanities**	**Math & Statistics**	**Science & Tech**
Collection		- | -	↑ | ↓	- | -	- | -	- | -
Reference Material		- | -	- | -	- | ↓	- | -	- | -
Simulation	- | ↓	↑ | -	- | ↓	- | -	- | -	↑ | ↓
Tutorial		↑ | ↓	↑ | ↓	↑ | -	- | -	- | ↓

Table 4: Tendencies of the accuracies according to the number of neurons used for training (good|not-good)

In table 4, the minus (-) represents no association between the accuracy of the model for classifying a given class and the number of neurons, the up arrow (↑) represents a positive association, and the down arrow (↓) a negative association. In each cell, the first signal corresponds to the tendency (positive or negative) of the association for classifying *good* LOs, and the second one to the tendency for classifying *not-good* LOs. As it can be seen in the table, the analysis shows associations and tendencies between the number of neurons and the accuracies for some classes of quality in some specific subsets.

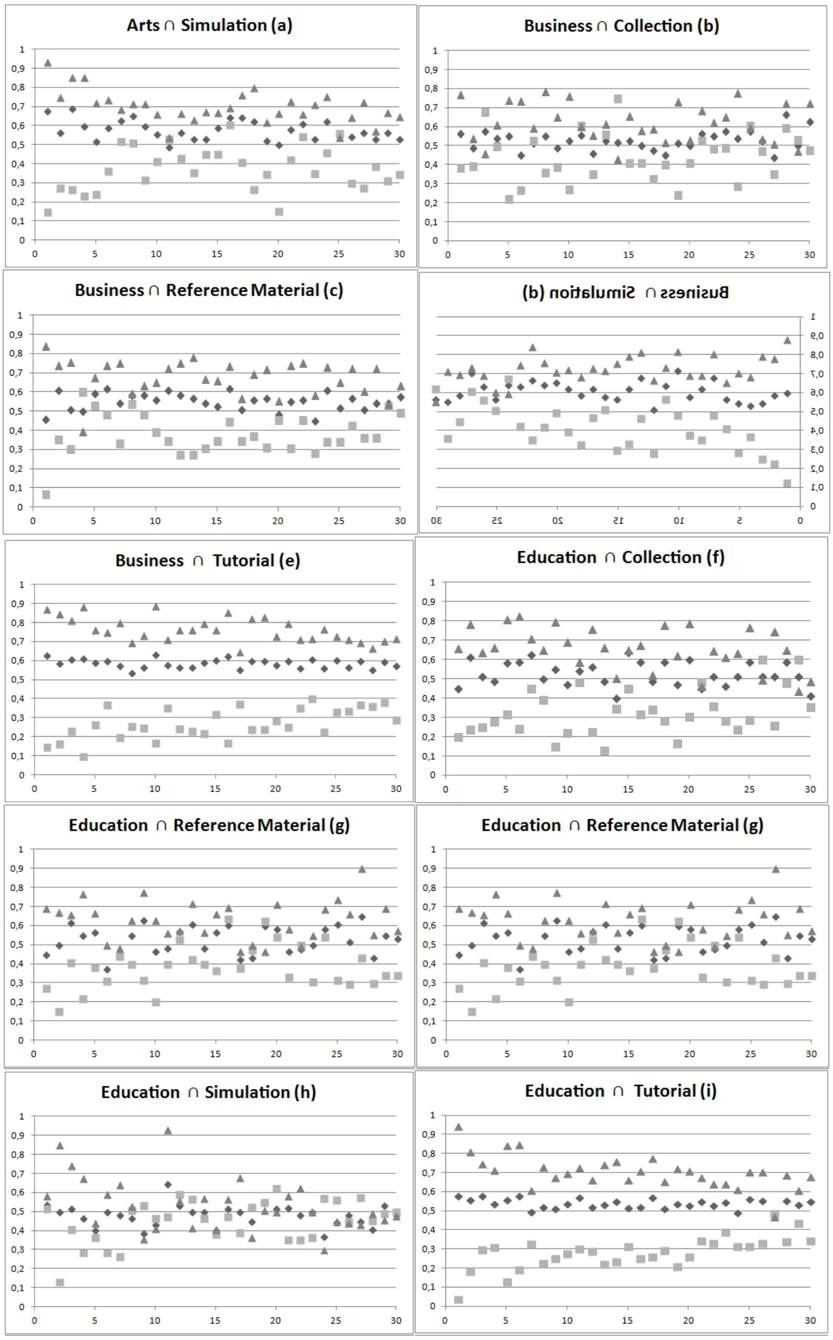
For instance, positive associations between the number of neurons and accuracies for the classification of *good* resources were found in the 6 (six) following subsets: *Business* ∩ *Tutorial*, *Business* ∩ *Simulation, Education* ∩ *Tutorial*, *Education* ∩ *Collection, Science & Technology* ∩ *Simulation* and *Humanities* ∩ *Tutorial.* Moreover, negative associations between the number of neurons and the accuracies for the classification of *not-good* LOs were found in 8 (eight) following subsets: *Business* ∩ *Tutorial*, *Arts* ∩ *Simulation*, *Education* ∩ *Simulation*, *Education* ∩ *Collection*, *Education* ∩ *Humanities*, *Education* ∩ *Tutorial*, *Science & Technology* ∩ *Tutorial* and *Science & Technology* ∩ *Simulation.* Finally, we found no positive associations between the number of neurons and the accuracies for the classification of *not-good* LOs; neither negative associations between the number of neurons and the accuracies for the classification of *good* LOs.

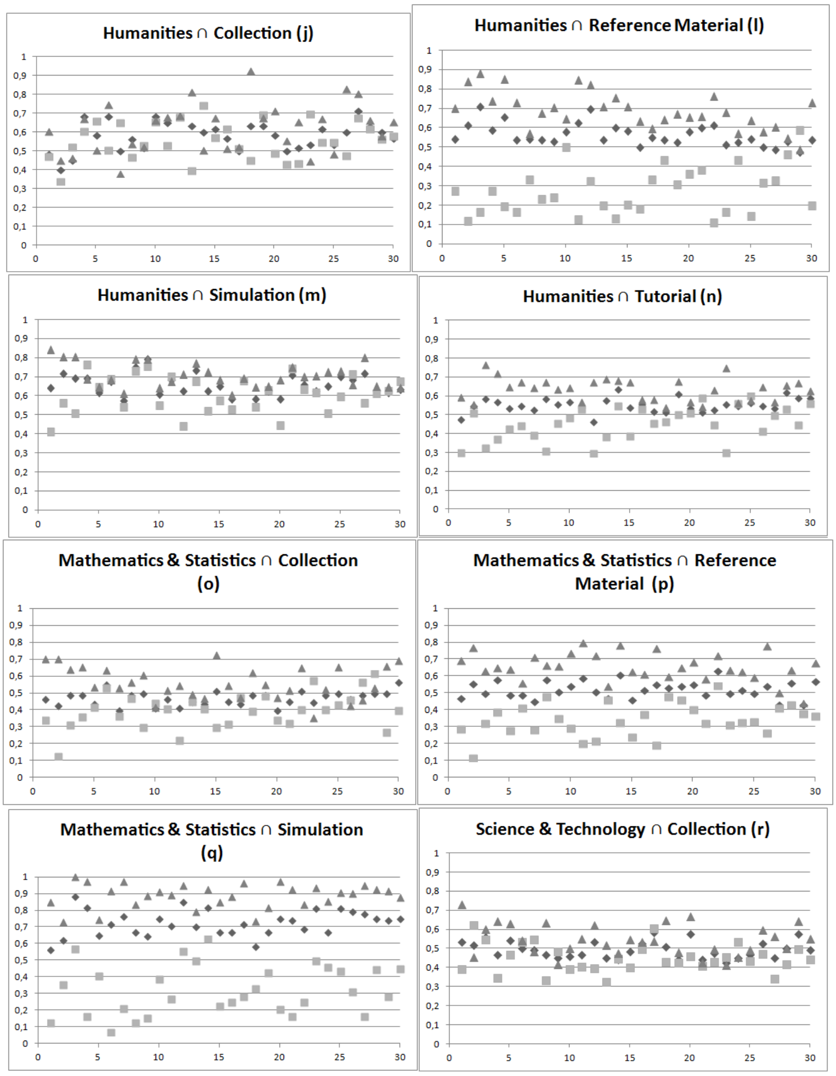
In order to evaluate how to select the best models for quality assessment, it is necessary to understand the behaviour of the models for classifying both classes of quality included in the datasets. Considering that, a Spearman's rank correlation (r_s) analysis was also carried out to evaluate whether there are associations between the accuracies of the models for classifying *good* and *not-good* resources. Such analysis serves to evaluate the trade-offs of selecting or not a given model for the present purpose. The results of this analysis are shown in Table 5.

Subset	**Arts**	**Business**	**Education**	**Humanities**	**Math & Statistics**	**Science & Tech**
Collection		-0,47	-0,52	N	-0,53	N
Reference Material		-0,53	N	-0,72	-0,51	-0,56
Simulation	-0,42	N	-0,53	N	N	-0,74
Tutorial		-0,87	-0,78	-0,6	-0,47	-0,56

Table 5: Spearman's rank correlation (rs) between accuracies of the models

Table 5 presents the correlations between the accuracies of models for classifying *good* and *not-good* resources and considering a 95% level of significance. N stands for no significant correlation. According to the values shown in the table, most of models presented strong negative associations between accuracies for classifying *good* and *not-good* LOs. Considering the findings of both analyses, it is recommended to take into account that accuracy for the classification of one class of LO increases at the same time that accuracy for the classification of the other class decreases. It is necessary then to establish the desirable cut-off point for accuracies so that models can be applied in our problem.


Arts ∩ Simulation (a)
Business ∩ Collection (b)
Business ∩ Reference Material (c)
Business ∩ Simulation (d)
Business ∩ Tutorial (e)
Education ∩ Collection (f)
Education ∩ Reference Material (g)
Education ∩ Reference Material (g)
Education ∩ Simulation (h)
Education ∩ Tutorial (i)
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1
0
5
10
15
20
25
30


Humanities ∩ Collection (j)

Humanities ∩ Reference Material (l)

Humanities ∩ Simulation (m)

Humanities ∩ Tutorial (n)

Mathematics & Statistics ∩ Collection (o)

Mathematics & Statistics ∩ Reference Material (p)

Mathematics & Statistics ∩ Simulation (q)

Science & Technology ∩ Collection (r)

Figure 2: Accuracies of the models versus number of neurons.

In this paper, we are dealing with a two-class problem. This means to say that new LOs must be classified between *good* or *not-good*, and the correct classification by a merely random decision is of 50%. Therefore, we established accuracies greater than 50% for the simultaneous classification of *good* and *not-good* LOs as the minimum for a model to be considered valuable. Table 6 presents the overall accuracies, and the accuracies for classifying *good* and *not-good* LOs for the 3 finest models for each subset. The models are ordered by the accuracy for classifying *good* LOs.

Subset	N	O	G	NG	Subset	N	O	G	NG
Arts ∩ Simulation	16	0.65	0.61	0.70	Business ∩ Collection	11	0.56	0.61	0.60
	25	0.55	0.56	0.54		25	0.57	0.60	0.59
	22	0.61	0.54	0.66		28	0.66	0.59	0.72
Business ∩ Reference	8	0.58	0.54	0.59	Business ∩ Simulation	24	0.64	0.67	0.60
	5	0.59	0.53	0.68		30	0.57	0.62	0.55
	29	0.54	0.53	0.53		27	0.70	0.61	0.73
Business ∩ Tutorial	23	0.61	0.40	0.72	Education ∩ Collection	26	0.51	0.6	0.49
	29	0.59	0.38	0.71		29	0.51	0.6	0.44
	17	0.55	0.37	0.65		21	0.45	0.48	0.46
Education ∩ Reference	16	0.60	0.63	0.70	Education ∩ Simulation	20	0.52	0.62	0.5
	20	0.58	0.54	0.71		12	0.53	0.59	0.56
	24	0.58	0.54	0.69		19	0.55	0.55	0.51
Education ∩ Tutorial	27	0.47	0.49	0.47	Humanities ∩ Collection	14	0.6	0.75	0.51
	29	0.53	0.43	0.61		19	0.63	0.69	0.68
	23	0.54	0.39	0.64		27	0.72	0.68	0.80
Humanities ∩ Reference Material	29	0.47	0.59	0.49	Humanities ∩ Simulation	4	0.69	0.76	0.69
	10	0.58	0.5	0.65		9	0.79	0.75	0.79
	28	0.53	0.46	0.55		21	0.71	0.74	0.75
Humanities ∩ Tutorial	25	0.56	0.60	0.58	Math & Statistics ∩ Collection	28	0.5	0.61	0.54
	21	0.51	0.59	0.54		27	0.49	0.57	0.46
	30	0.59	0.56	0.63		6	0.55	0.53	0.64
Math ∩ Reference Material	22	0.63	0.54	0.72	Math & Statistics ∩ Simulation	14	0.81	0.63	0.93
	18	0.53	0.48	0.60		3	0.88	0.57	1
	8	0.58	0.48	0.67		12	0.85	0.56	0.95
Math ∩ Tutorial	26	0.69	0.79	0.64	Science & Tech ∩ Collection	17	0.58	0.60	0.54
	25	0.70	0.77	0.61		3	0.56	0.54	0.60
	9	0.64	0.75	0.63		6	0.50	0.53	0.54
Science & Tech ∩ Reference Material	19	0.59	0.63	0.56	Science & Tech ∩ Simulation	29	0.57	0.58	0.61
	16	0.55	0.58	0.58		19	0.58	0.52	0.62
	20	0.53	0.54	0.52		16	0.58	0.50	0.62
Science & Tech ∩ Tutorial	28	0.64	0.50	0.72					
	14	0.56	0.45	0.61					
	17	0.56	0.44	0.65					

Table 6: Top 3 models for each subset (ordered by the accuracies for the classification of good LOs)

In table 6, the number of neurons is presented in column N, overall accuracies in column O, precisions for the classification of *good* LOs in column G, and precisions for the classification of *not-good* LOs in column NG. As it is shown in the table, only 14 of the 63 models did not present the cut-off accuracies established for the study (white cells in the table). Besides, 33 models presented both accuracies between 50% and 59.90% (gray cells in the table), and 12 models presented both accuracies greater than 60% (black cells in the table). It was also possible to find 2 (two) models where both accuracies were greater than 70% (for the *Humanities ∩ Simulation* subset). Models did not present the minimum cut-off accuracies to the following three subsets *Business ∩ Tutorial*, *Education ∩ Collection* and *Education ∩ Tutorial*. At last, the best results were found for the following subsets: *Humanities ∩ Simulation*, *Mathematics ∩ Tutorial*, *Humanities ∩ Collection*, *Business ∩ Simulation*, *Arts ∩ Simulation*, and *Business ∩ Collection*. It is difficult to state the reasons why was not possible to create acceptable models for all subsets, but this may be because potential features associated to quality on those subsets were not collected by the crawler.

Considering that classification provided by the models will be used as information during the ranking process, it is important to evaluate the shortcomings of low accuracies for classifying *good* LOs in comparison to low accuracies for classifying the *not-good* ones. In the case that *good* LOs are misclassified as *not-good*, such materials would just be put in bad ranked positions, which would be equivalent to the situation of not using the models. On the contrary, if *not-good* LOs are misclassified as *good*, this would increase the chances of *not-good* LOs be accessed by the users, thus misleading the repository audience and putting in discredit the ranking mechanism.

5 Extending the methodology to other repositories – the case of Connexions

Among the most recent LORs initiatives, Connexions occupies a highlight position that can be observed by its exponential growth of contributors and materials in the last few years. The main conception of this repository is to allow users to create and share learning materials in a collaborative way. According to [Ochoa, 2010], the main reason for Connexions success lies on the fact that the repository implements a new paradigm where materials are created through social interactions by the members of community. Materials in Connexions can be developed in the form of modules (the most granular piece of knowledge) and collections (groups of modules structured into course notes) and can be used, reused, assembled and shared under a Creative Commons license.

Connexions implements quality assessment through the use of endorsements given by organizations (universities, industries, companies) and individuals [Kelty, Burrus, & Baraniuk, 2008]. They called these evaluations *lenses.* In Connexions, LOs acquire higher value as they accumulate more *lenses* from others. Besides the use of endorsements (*lenses*), Connexions also ranks LOs based on some popularity measures such as the number of accesses, and on the ratings given by the community of users. For the present experiment, we used the number of lenses of a given LO to generate the classes of quality.

5.1 Methodology

The methodology followed here is the same as the one described in Section 3. However, some considerations are important to be made. LOs in Connexions are developed inside their own platform thus presenting a more uniform structure then materials from MERLOT. Considering that, LOs in Connexions are not differentiated by their type, as they all present the same structure of modules. Besides, Connexions is essentially composed by contents in the form of text books, and because of that, we gathered only 12 the metrics that are supposed to appear in such kind of material. At last, due to a limitation of our crawler, the metrics were evaluated considering the complete sample and without considering the possible categorisation of the resources into different disciplines. Metrics in italic in Table 1 are the ones collected for this experiment.

5.2 Data collection

We collected information form a total of 8,855 LOs. The lenses distribution in Connexions is significantly different from the distribution of ratings in MERLOT (see fig. 3). The amount of LOs with one or more lenses is very close to the amount of LOs without lenses. Precisely, 53.55% (4,742) of the LOs in Connexions are endorsed at least one time, 1.84% (163) of the LOs are endorsed from 2 to 4 times, and 44.46% (3,950) of the LOs are not endorsed.

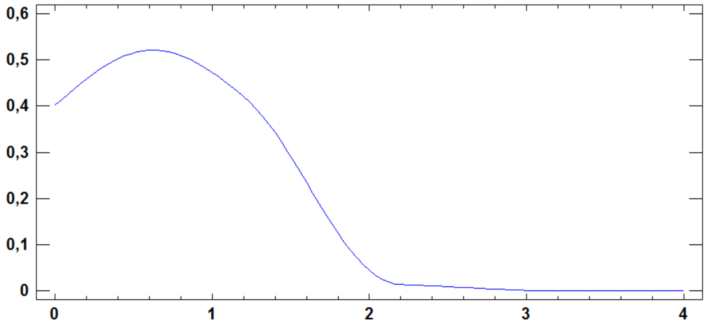


Figure 3: Distribution of the number of lenses per LO

LOs without endorsements were excluded from this study, and the remaining 4,905 were split up into two different classes of quality: 1) those with just one endorsement, and 2) those with two or more endorsements. These classes and their sizes are presented in table 8.

Class	Amount	Percentage of the Sample
One endorsement	4,742	96.67
Two or more endorsements	163	03.33
Total	4,905	100.00

Table 8: Classes of quality

5.3 About the models

After running the data mining algorithm, we were able to find models for classifying resources in Connexions between those with *one endorsement* and those with *two or more endorsements*. The generated models presented a different behaviour of those created for MERLOT in terms of the correlations between the number of neurons and the accuracies for the classification of the output classes. In here, the number of neurons is positively correlated (at a 95% level of significance) with both accuracies (for classifying resources with *one endorsement* and with *two or more endorsements*). Precisely, the number of neurons is moderately correlated with the accuracies for classifying resources with *one endorsement* ($r_s = 0.42$) and strongly correlated with the accuracies for classifying resources with *two or more endorsements* ($r_s = 0.63$). As it can be seen in figure 4, the models presented very high accuracies for classifying resources with *one endorsement* (99%) and lower accuracies for classifying resources with *two or more endorsements*, a similar situation.

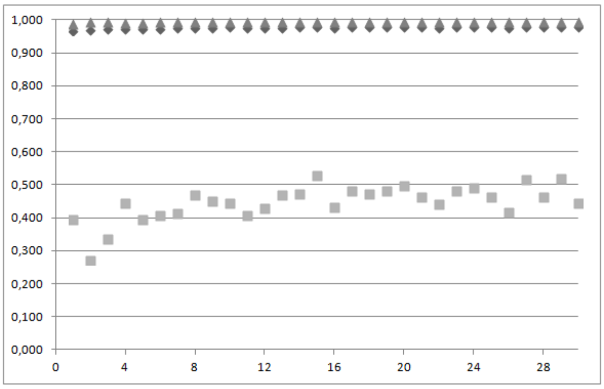


Figure 4: Accuracies of the models versus number of neurons.

In table 9 we present the top 10 models ordered by the accuracies for classifying resources with *two or more endorsements*. As we can see, it was possible to generate 4 models that fit the same rule we applied earlier to the models for MERLOT (accuracies greater than 50% for the correct classification of both output classes). However, in here, the models for classifying resources with *two or more endorsements* (correspondent to the *not-good* resources class in MERLOT) presented their highest accuracies much lower than those from the models generated for MERLOT. This is probably a consequence of the small size of this group in the sample (only 3.33% of the total sample). While in the experiment with MERLOT we were able to split the samples more equally (by using the terciles of the existing ratings), in here it was not possible to do the same and thus there is a huge concentration of resources in one of the output classes.

Neurons	Overall	One endorsement	Two or more endorsements
15	0.98	0.99	0.53
29	0.98	0.99	0.52
27	0.98	1.00	0.52
20	0.98	0.99	0.50
24	0.98	1.00	0.49
17	0.98	0.99	0.48
23	0.98	0.99	0.48
19	0.98	0.99	0.48
14	0.98	0.99	0.47
18	0.98	0.99	0.47

Table 9: Top 10 models (ordered by the accuracies for classifying resources with two-or-more-endorsements)

6 Conclusions

Learning Object Repositories commonly consider quality information of their resources as a relevant feature to rank search results. This kind of evaluative information is primarily generated by the community of users of such platforms but are not available for all the LOs catalogued in the repositories. This is mainly because LOs inside LORs tend to increase more quickly than the capacity of the community to evaluate them. Therefore, many LOs that could occupy better positions during the search and retrieval process, remain unused until someone decides to evaluate them. The present work has presented the feasibility of an alternative approach to automatically assess quality of LOs based on features that can be automatically extracted from the LOs themselves. We tested our methodology with data collected from two distinct repositories, namely: MERLOT and Connexions.

For the experiments performed with MERLOT datasets, we used the ratings given by the community of experts as the reference quality information to be predicted by the models. Among other results, it is worth mentioning a model for *Humanities* ∩ *Simulation* which was able to classify *good* LOs with 75% of precision, and *not-good* LOs with 79%; along with a model trained for *Mathematics* ∩ *Tutorial* that reached precisions of 79% and 64% to classify *good* and *not-good* LOs respectively. In the experiment with data from Connexions all models presented very high accuracies for the classification of resources *with one endorsement* and we have found 4 models able to classify resources *with two or more endorsements* with accuracies varying from 50% to 53%.

Models created in this work could be used to estimate quality for those LOs still not assessed by members of the community, therefore helping to optimize the results provided by search engines in repositories. As the models would be implemented inside LORs and classifications would serve just as input information for search engines, it is not necessarily required that models provide explanations about their reasoning. Black box models (such as ANNs) can perfectly be used in such a scenario.

Resources recently added to LORs would be decidedly benefited by the use of such models, once they hardly receive assessments just after their inclusion. Once the LOs receive formal evaluations from the community of the repository, these artificial ratings provided by the models could be ignored. Moreover, this new arrived evaluation could be used as feedback information to evaluate the efficiency of the models and check whether users agree with models classifications. It is noteworthy to mention that even though such quality estimation could be internally useful for the search and retrieval mechanisms of the repository, it should not be displayed to the final user given its uncertain reliability degree.

7 Future work

In this work, we evaluated accuracies for predictive models created by the proposed approach. In future works, increasing the amount of performance measures can improve the quality of these predictive models. [Cichosz, 2011] presents a set of potentially usable performance measures that could help on that. Moreover, other metrics computed from LO content could be included, such as the number of colours, font styles, and links (redundant and broken), the presence of advertisements, and readability measures. In the present work we used ratings from the community of experts in MERLOT as the basis to determine the quality of the resources. Models could also be tested by using the ratings given by the community of users. Interesting insights could emerge from the comparison of models generated by using ratings from these two different communities. Moreover, the inclusion of the granularity of the learning objects as one of the metrics is also another possible direction for future work. At last, experiments comparing the rankings obtained by current search and retrieval mechanisms of MERLOT and Connexions against rankings obtained by using quality information generated by the predictive models, could also help to better evaluate the efficiency of such an approach.

Another viable future work is the development of white box models which are easily understandable and interpretable (e.g. decision trees, rules-based models) [Cechinel, Silva Camargo, Sánchez-Alonso, & Sicilia, 2012]. Such models could help to explain the reasoning behind the final quality classification and thus provide useful information for LO developers who need an introductory assessment of their materials. Information presented by these models can point out weak and strong aspects of each LO, and lead developers to change their materials before publishing them inside LORs. LORs could provide access to such models as a service where LO creators can consult and check which intrinsic features of the resource are influencing their quality. Previous work has already stated the importance of tools to support the creation of Learning Objects [Dodero, Díaz, Aedo, & Cabezuelo, 2005].

At last, the present approach could be used as a complement to the traditional assessment approaches currently in use, so that preliminary information about learning objects quality is provided before performing a more time-consuming assessment.

Acknowledgements

This research has been funded by FAPERGS (Foundation for Research Support of the State of Rio Grande do Sul - Brazil) through the Pesquisador Gaúcho Program - PQG - Edital 02/2014, and by CYTED (Ibero-American Programme for Science, Technology and Development) as part of project "RIURE - Ibero-American Network for the Usability of Learning Repositories" (code 513RT0471).

References

[Bethard, S., Wetzer, P., Butcher, K., Martin, J. H., & Sumner, T., 2009] Bethard, S., Wetzer, P., Butcher, K., Martin, J. H., & Sumner, T. (2009). *Automatically characterizing resource quality for educational digital libraries*. Paper presented at the Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries, Austin.

[Bishop, C. M., 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*: Springer.

[Blumenstock, J. E., 2008] Blumenstock, J. E. (2008). *Size matters: word count as a measure of quality on wikipedia*. Paper presented at the Proceedings of the 17th international conference on World Wide Web, Beijing, China.

[Cafolla, R., 2006] Cafolla, R. (2006). Project MERLOT: Bringing Peer Review to Web-Based Educational Resources. *Journal of Technology and Teacher Education, 14*(2), 313-323.

[Cechinel, C., 2012] Cechinel, C. (2012). *Empirical Foundations for Automated Quality Assessment of Learning Objects inside Repositories.* (Ph.D. Doctoral Thesis), University of Alcalá, Alcalá de Henares.

[Cechinel, C., Camargo, S. d. S., Ochoa, X., Sánchez-Alonso, S., & Sicilia, M.-Á., 2012] Cechinel, C., Camargo, S. d. S., Ochoa, X., Sánchez-Alonso, S., & Sicilia, M.-Á. (2012). *Populating Learning Object Repositories with Hidden Internal Quality Information.* Paper presented at the Recommender Systems in Technology Enhanced Learning, Saarbrücken, Germany, .

[Cechinel, C., Sánchez-Alonso, S., & García-Barriocanal, E., 2011] Cechinel, C., Sánchez-Alonso, S., & García-Barriocanal, E. (2011). Statistical profiles of highly-rated learning objects. *Computers & Education, 57*(1), 1255-1269. doi: 10.1016/j.compedu.2011.01.012

[Cechinel, C., Sánchez-Alonso, S., & Sicilia, M.-Á., 2009] Cechinel, C., Sánchez-Alonso, S., & Sicilia, M.-Á. (2009). Empirical Analysis of Errors on Human-Generated Learning Objects Metadata. In F. Sartori, M. Á. Sicilia & N. Manouselis (Eds.), *Metadata and Semantic Research* (Vol. 46, pp. 60-70): Springer Berlin Heidelberg.

[Cechinel, C., Silva Camargo, S., Sánchez-Alonso, S., & Sicilia, M.-Á., 2012] Cechinel, C., Silva Camargo, S., Sánchez-Alonso, S., & Sicilia, M.-Á. (2012). On the Search for Intrinsic Quality Metrics of Learning Objects. In J. Dodero, M. Palomo-Duarte & P. Karampiperis (Eds.), *Metadata and Semantics Research* (pp. 49-60): Springer Berlin Heidelberg.

[Cichosz, P., 2011] Cichosz, P. (2011). Assessing the quality of classification models: Performance measures and evaluation procedures. *Central European Journal of Engineering, 1*(2), 132-158. doi: 10.2478/s13531-011-0022-9

[Dalip, D. H., Gonçalves, M. A., Cristo, M., & Calado, P., 2011] Dalip, D. H., Gonçalves, M. A., Cristo, M., & Calado, P. (2011). Automatic Assessment of Document Quality in Web

Collaborative Digital Libraries. *J. Data and Information Quality, 2*(3), 1-30. doi: 10.1145/2063504.2063507

[Dodero, J. M., Díaz, P., Aedo, I., & Cabezuelo, A. S., 2005] Dodero, J. M., Díaz, P., Aedo, I., & Cabezuelo, A. S. (2005). Integrating Ontologies into the Collaborative Authoring of Learning Objects. *J. UCS, 11*(9), 1568-1578.

[Hagan, M. T., & Menhaj, M. B., 1994] Hagan, M. T., & Menhaj, M. B. (1994). Training feedforward networks with the Marquardt algorithm. *Neural Networks, IEEE Transactions on, 5*(6), 989-993. doi: 10.1109/72.329697

[Han, P., Kortemeyer, G., Krämer, B., & von Prümmer, C., 2008] Han, P., Kortemeyer, G., Krämer, B., & von Prümmer, C. (2008). Exposure and Support of Latent Social Networks among Learning Object Repository Users. *Journal of the Universal Computer Science, 14*(10), 1717-1738. doi: citeulike-article-id:3558788

[Ivory, M. Y., & Hearst, M. A., 2002] Ivory, M. Y., & Hearst, M. A. (2002). *Statistical profiles of highly-rated web sites*. Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves, Minneapolis, Minnesota, USA.

[Kelty, C. M., Burrus, C. S., & Baraniuk, R. G., 2008] Kelty, C. M., Burrus, C. S., & Baraniuk, R. G. (2008). Peer Review Anew: Three Principles and a Case Study in Postpublication Quality Assurance. *Proceedings of the IEEE, 96*(6), 1000-1011.

[Kewen, W., Qinghua, Z., Yuxiang, Z., & Hua, Z., 2010] Kewen, W., Qinghua, Z., Yuxiang, Z., & Hua, Z. (2010, 7-8 Aug. 2010). *Mining the Factors Affecting the Quality of Wikipedia Articles.* Paper presented at the International Conference of Information Science and Management Engineering (ISME).

[Koukourikos, A., Stoitsis, G., & Karampiperis, P., 2012] Koukourikos, A., Stoitsis, G., & Karampiperis, P. (2012). *Sentiment Analysis: A tool for Rating Attribution to Content in Recommender Systems* Paper presented at the Recommender Systems in Technology Enhanced Learning, Saarbrücken, Germany, .

[Mendes, E., Hall, W., & Harrison, R., 1998] Mendes, E., Hall, W., & Harrison, R. (1998). Applying Metrics to the Evaluation of Educational Hypermedia Applications. *Journal of Universal Computer Science, 4*(4), 382-403. doi: 10.3217/jucs-004-04-0382

[Nesbit, J. C., Belfer, K., & Leacock, T., 2003] Nesbit, J. C., Belfer, K., & Leacock, T. (2003). Learning object review instrument (LORI). E-learning research and assessment network. Retrieved from http://www.elera.net/eLera/Home/Articles/LORI%20manual.

[Ochoa, X., 2010] Ochoa, X. (2010). Connexions: a Social and Successful Anomaly among Learning Object Repositories. *Journal of Emerging Technologies in Web Intelligence, 2*(1).

[Ochoa, X., & Duval, E., 2008] Ochoa, X., & Duval, E. (2008). Relevance Ranking Metrics for Learning Objects. *Learning Technologies, IEEE Transactions on, 1*(1), 34-48. doi: http://dx.doi.org/10.1109/TLT.2008.1

[Ochoa, X., & Duval, E., 2009] Ochoa, X., & Duval, E. (2009). Quantitative Analysis of Learning Object Repositories. *Learning Technologies, IEEE Transactions on, 2*(3), 226-238.

[Santos, H. L. d., & Cechinel, C., 2015] Santos, H. L. d., & Cechinel, C. (2015). *Geração automática de avaliações de objetos de aprendizagem por meio de mineração de textos.* Paper presented at the Anais dos Workshops do Congresso Brasileiro de Informática na Educação.

[Sicilia, M.-Á., García-Barriocanal, E., Pages, C., Martinez, J. J., & Gutierrez, J. M., 2005] Sicilia, M.-Á., García-Barriocanal, E., Pages, C., Martinez, J. J., & Gutierrez, J. M. (2005).

Complete metadata records in learning object repositories: some evidence and requirements. *International Journal of Learning Technology (IJLT), 1*(4), 411-424. doi: http://dx.doi.org/10.1504/IJLT.2005.007152

[Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C., 2007] Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A framework for information quality assessment. *J. Am. Soc. Inf. Sci. Technol., 58*(12), 1720-1733. doi: 10.1002/asi.v58:12

[Stvilia, B., Twidale, M. B., Smith, L. C., & Gasser, L., 2005] Stvilia, B., Twidale, M. B., Smith, L. C., & Gasser, L. (2005). *Assessing information quality of a community-based encyclopedia.* Paper presented at the Proceedings of the International Conference on Information Quality - ICIQ 2005.

[Vuorikari, R., Manouselis, N., & Duval, E., 2008] Vuorikari, R., Manouselis, N., & Duval, E. (2008). Using Metadata for Storing, Sharing and Reusing Evaluations for Social Recommendations: the Case of Learning Resources. *Social Information Retrieval Systems: Emerging Technologies and Applications for Searching the Web Effectively. Hershey, PA: Idea Group Publishing*, 87–107.

[Xu, L., Hoos, H. H., & Leyton-Brown, K., 2007] Xu, L., Hoos, H. H., & Leyton-Brown, K. (2007). *Hierarchical hardness models for SAT.* Paper presented at the Proceedings of the 13th international conference on Principles and practice of constraint programming, Providence, RI, USA.