Towards Automated Evaluation of Learning Resources Inside Repositories

Cristian Cechinel, Sandro da Silva Camargo, Salvador Sánchez-Alonso, and Miguel-Ángel Sicilia

Abstract It is known that current Learning Object Repositories adopt strategies for quality assessment of their resources that rely on the impressions of quality given by the members of the repository community. Although this strategy can be considered effective at some extent, the number of resources inside repositories tends to increase more rapidly than the number of evaluations given by this community, thus leaving several resources of the repository without any quality assessment. The present work describes the results of two experiments to automatically generate quality information about learning resources based on their intrinsic features as well as on evaluative metadata (ratings) available about them in MERLOT repository. Preliminary results point out the feasibility of achieving such goal which suggests that this method can be used as a starting point for the pursuit of automatically generation of internal quality information about resources inside repositories.

Keywords Learning repositories • Quality metrics • Automatic assessment

C. Cechinel (✉)
Distance Learning Center, Federal University of Pelotas,
Felix da Cunha, 630 Centro, Pelotas, RS, Brazil
e-mail: contato@cristiancechinel.pro.br

S.da Silva Camargo
Computer Engineering Course, Federal University of Pampa,
Caixa Postal 07, 96400-970 Bagé, RS, Brazil
e-mail: camargo.sandro@gmail.com

S. Sánchez-Alonso • M.-Á. Sicilia
Information Engineering Research Unit, Computer Science Department, University of Alcalá, Ctra. Barcelona km. 33.6, 28871 Alcalá de Henares, Madrid, Spain
e-mail: salvador.sanchez@uah.es; msicilia@uah.es

N. Manouselis et al. (eds.), *Recommender Systems for Technology Enhanced Learning: Research Trends and Applications*, DOI 10.1007/978-1-4939-0530-0_2,

Introduction

Current Learning Object Repositories (LORs) normally adopt strategies for the establishment of quality of their resources that rely on the impressions of usage and evaluations given by the members of the repository community (ratings, tags, comments, likes, lenses). All this information together constitute a collective body of knowledge that further serves as an external memory that can help other individuals to find resources according to their needs. Inside LORs, this kind of evaluative metadata [1] is also used by search and retrieval mechanisms for properly ranking and recommending resources to the community of users of the repository.

Although such strategies can be considered effective at some extent, the amount of resources inside repositories is rapidly growing every day [2] and it becomes impractical to rely only on human effort for such a task. For instance, on a quick look at the summary of MERLOT's recent activities, it is possible to observe that in a short period of 1 month (from May 21 to June 21, 2011), the amount of new resources catalogued in the repository was nine times more than the amount of new ratings given by experts (peer-reviewers), six times more than the amount of new comments (and users ratings) and three times more than the amount of new bookmarks (personal collections). This situation of leaving many resources of the current repositories without any measure of quality at all (and consequently unable or at least on a very disadvantaged position to compete for a good position during the process of search and retrieval) has raised the concern for the development of new automated techniques and tools that could be used to complement existing manual approaches. On that direction, Ochoa and Duval [3] developed a set of metrics for ranking repository search results according to three dimensions of relevance (topical, personal and situational) and by using information obtained from the learning objects metadata, from the user queries, and from other external sources such as the records of historical usage of the resources. This authors contrasted the performance of their approach against the text-based ranking traditional methods and have found significant improvements in the final ranking results. Moreover, Sanz-Rodriguez et al. [4] proposed to integrate several distinct quality indicators of learning objects of MERLOT along with their usage information into one overall quality indicator that can be used to facilitate the ranking of learning objects.

These mentioned approaches for automatically measuring quality (or calculating relevance) according to specific dimensions depend either on the existence and availability of metadata attached to the resources (or inside the repositories), or on measures of popularity about the resources that are obtained only when the resource is publicly available after a certain period of time. As metadata may be incomplete/inaccurate [5] and these measures of popularity will be available just for "old" resources, we propose to apply an alternative approach for this problem. The main idea is to identify intrinsic measures of the resources (i.e., features that can be calculated directly from the resources) that are associated to quality and that can be used in the process of creating models for automated quality assessment.

In fact, this approach was recently tested by Cechinel et al. [6] who developed highly-rated profiles of learning objects available in MERLOT, and have generated Linear Discriminant Analysis (LDA) models based on 13 learning objects intrinsic features. The generated models were able to classify resources between good and not-good with 72.16 % of precision, and between good and poor with 91.49 % of precision. Among other things, these authors concluded that highly-rated learning objects profiles should be developed taking into consideration the many possible intersections among the different disciplines and types of materials available in MERLOT, as well as the group of evaluators who rated the resources (whether they are formed by experts or by the community of users). For instance, the mentioned models were created for materials of *Simulation* type belonging to the discipline of *Science & Technology*, and considering the perspective of the peer-reviewers ratings.

The present chapter reviews two experiments conducted towards the creation of models for automated quality assessment of learning resources inside MERLOT and that expand the previous work developed by Cechinel et al. [6]. The first experiment explores the creation of statistical profiles of highly-rated learning objects by contrasting information from *good* and *not-good* resources of three subsets of MERLOT repository and by using these profiles to generate models for quality assessment. The second experiment tests a slightly different and more algorithmic approach, i.e., the models are generated exclusively through the use of data mining algorithms. In this second experiment we also worked with a larger collection of resources and a considerably higher number of MERLOT subsets.

The rest of this chapter is structured as follows. "Background" presents existing research focused on identifying intrinsic quality features of resources. "Data Collection" describes the data collected for the experiments. "First Experiment: Statistical profiles of highly-rated resources" and "Second experiment: Algorithmic Approach" present the experiments and some discussion about the results on the generation and evaluation of automated models for quality assessment. Finally, conclusions and outlook are provided in "Conclusions and Outlook".

Background

Apart from the recent works by Cechinel et al. [6, 7], there is still no empirical evidence of intrinsic metrics that could serve as indicators of quality for LOs. However, there are some works in adjacent fields which can serve us as a source of inspiration. For instance, empirical evidence of relations from intrinsic information and other characteristics of LOs have been found in [8], where the authors developed a model for classifying the didactic functions of a learning object based on measures about the length of the text, the presence of interactivity and information contained in the HTML code (lists, forms, input elements). Mendes et al. [9] have identified evidence in some measures to evaluate sustainability and reusability of educational hypermedia applications, such as, the type of link and the structure and size of the application. Blumenstock [10] has found the length of an article (measured in

words) as a predictor of quality in Wikipedia. Moreover, Stvilia et al. [11] have been able to automatically discriminate high quality articles voted by the community of users from the rest of the articles of the collection. In order to do that, the authors developed profiles by contrasting metrics of articles featured as best articles by Wikipedia editors against a random set. The metrics were based on measures of the article edit history (total number of edits, number of anonymous user edits, for instance) and on the article attributes and surface features (number of internal broken links, number of internal links, number of images, for instance). At last, in the field of usability, Ivory and Hearst [12] have found that good websites contain (for instance) more words and links than the regular and bad ones.

Our approach is initially related exclusively to those aspects of learning objects that are displayed to the users and that are normally associated to the dimensions of presentation design and interaction usability included in LORI [13] and the dimension of information quality (normally mentioned in the context of educational digital libraries). Precisely, the references for quality assurance used in here are the ratings given by the peer-reviewers (experts) of the repository.

Data Collection

Two databases were collected from MERLOT (2009 and 2010) through the use of a crawler that systematically traversed the pages and collected information related to 34 metrics of the resources. The decision of choosing MERLOT lays mainly on the fact that MERLOT has one of the largest amount of registered resources and users, and it implements a system for quality assurance that works with evaluations given by experts and users of the repository. Such system can serve as baseline for the creation of the learning object classes of quality. As MERLOT repository is mainly formed by learning resources in the form of websites, we evaluated intrinsic metrics that are supposed to appear in such technical type of material (i.e., link measures, text measures, graphic measures and site architecture measures). The metrics collected for this study (see Table 1) are the same as used by Cechinel et al. [6] and some of them have also been mentioned in other works which tackled the problem of assessing quality of resources (previously presented in "Background").

Given that the resources in MERLOT vary considerably in size, a limit of two levels of depth was established for the crawler, i.e., metrics were computed for the root node (level 0—the home-page of the resource), as well as for the pages linked by the root node (level 1), and for the pages linked by the pages of the level 1 (level 2[1]). As it is shown in Table 1, some of the metrics refer to the total sum of the occurrences of a given attribute considering the whole resource, and other metrics refer to the average of this sum considering the number of the pages computed.

[1] Although this limitation may affect the results, the process of collecting the information is extremely slow and such limitation was needed. In order to acquire the samples used in this study, the crawler kept running uninterruptedly for 2 (in 2009) and 4 (in 2010) full months.

Table 1 Metrics collected for the study

Class of measure	Metric
Link measures	Number of links, number of unique[a] links, number of internal links[b], 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[c]
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

[a]The term unique stands for "non-repeated"
[b]The term internal refers to those links which are located at some directory below the root site
[c]For these metrics the average was not computed or does not exist

For instance, an object composed by 3 pages and containing a total of 30 images will have a total number of images equals to 30, and an average number of images equals to 10 (=30/3).

Classes of Quality

As the peer-reviewers ratings tend to concentrate above the intermediary rating 3, classes of quality were created using the terciles of the ratings for each subset (ratings in MERLOT vary from 1 to 5). Resources with ratings below the first tercile are classified as *poor*, resources with ratings equal or higher the first tercile and lower than the second tercile are classified as *average*, and resources with ratings equal or higher the second tercile are classified as *good*. The classes of quality *average* and *poor* were then joined in another class called *not-good* and were used as the output reference for generating and testing models for automated quality assessment of the resources

First Experiment: Statistical Profiles of Highly-Rated Resources

The collected sample contained 6,470 learning resources classified into 7 different disciplines and 9 distinct types of material, thus totalizing 63 different classes of possible learning object profiles. From the total, 1,257 (19.43 %) had at least one peer review rating and formed the final data sample. We have selected resources from the three subsets with the highest number of occurrences to generate and evaluate models for automated quality assessment in the context of peer-reviews thresholds. The selected subsets are (amounts in parenthesis): *Simulation* ∩ *Science and Technology* (97), *Simulation* ∩ *Mathematics and Statistics* (83), and *Tutorial* ∩ *Science and Technology* (83).

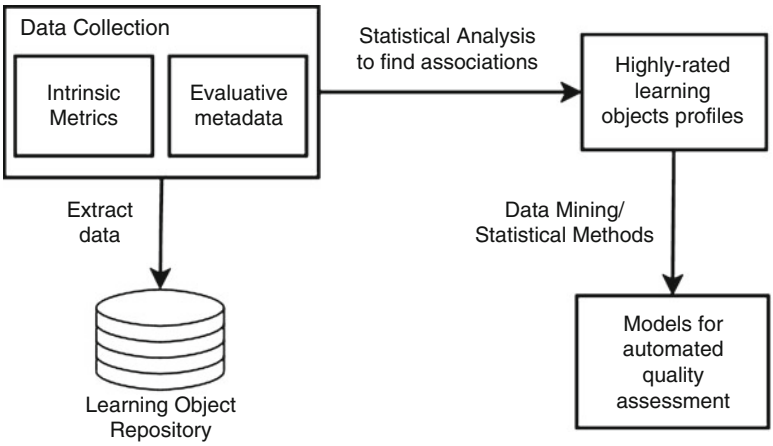

Fig. 1 Methodology for generating models for automated quality assessment

The methodology used for the present study was the development of highly-rated learning object profiles of MERLOT. The study described in this chapter is based on the methodology applied by Ivory and Hearst [12], as well as on the methodology described on García-Barriocanal and Sicilia [14] and Cechinel et al. [6]. The created profiles were then further used to generate models for automated quality assessment of learning objects. Figure 1 gives a general idea of the methodology applied here.

The analysis was conducted by contrasting intrinsic metrics from the groups between *good* and *not-good*[2] resources, and by observing if they presented significant differences between them. As the samples did not follow a normal distribution, a Mann-Whitney (Wilcoxon) test was performed to evaluate whether the classes presented differences between their medians, and a Kolmogorov-Smirnov test was applied to evaluate if the classes presented distinct distributions. When both distributions and medians presented significant differences, the metric was considered as a potential indicator of quality. The tendency of each metric (whether they influence negatively or positively the quality of the resource) was observed by comparing the median values of the samples. Table 2 presents the metrics that are associated to highly rated learning objects and their tendencies for each analyzed subset.

As it can be seen in Table 2, the metrics present different associations and tendencies depending on the given subsets. For instance, for the subset *Simulation* ∩ *Science and Technology*, seven metrics are positively associated to quality and six metrics negatively associated. On the other hand, for the subset of *Simulation* ∩ *Mathematics and Statistics* all metrics associated to quality present positive tendencies and for the subset of *Tutorial* ∩ *Science and Technology* all metrics associated to quality present negative tendencies.

[2] The so-called not-good group was formed by the union of the *average* group and the *poor* group.

Table 2 Significant discriminators and tendencies of the metrics for the good category of the selected subsets

Metric	Simulation ∩ science and technology	Simulation ∩ mathematics and statistics	Tutorial ∩ science and technology
Number of links	–	Y↑	Y↓
Number of unique links	–	Y↑	(Y)↓
Number of internal links	–	(Y)↑	Y↓
Number of unique internal links	–	(Y)↑	(Y)↓
Number of external links	Y↓	–	(Y)↓
Number of unique external links	Y↓	–	–
Size of the page (in bytes)		Y↑	(Y)↓
Number of images	(Y)↑	Y↑	–
Total size of the images (in bytes)	Y↑	Y↑	–
Number of scripts	Y↑	Y↑	–
Number of words	–	–	(Y)↓
Number of words that are links	–	–	Y↓
Number of applets	Y↓	–	–
Average number of unique internal links	–	–	(Y)↓
Average number of internal links	–	–	Y↓
Average number of unique external links	Y↓	–	–
Average number of external links	Y↓	–	(Y)↓
Average number of unique links	–	(Y)↑	Y↓
Average number of links	–	–	Y↓
Average number of applets	Y↓	–	–
Average number of images	Y↑	–	–
Average size of the pages	Y↑	–	–
Average size of the images	Y↑	Y↑	–
Average number of scripts	Y↑	(Y)↑	–
Total	13	11	13

Note: Y stands for both differences (medians and distributions) at the same time. The overall analysis was conducted for a 95 % confidence level; information in parenthesis means the results are significant at the 90 % level. Moreover (↑) stands for a positive contribution and (↓) stands for negative contribution

The Models

We created models for automated quality assessment of the resources through Data Mining Classification Algorithms (DMCA). Classification algorithms aim to construct models capable of associating each record of a given dataset to a labeled category. We have used WEKA [15] to generate and test models for the classification of resources between *good* and *not-good*, and among *good*, *average* and *poor* resources through the following classification algorithms: J48, SimpleCart, PART, Multilayer Perceptron Neural Network and Bayesian Network. Tables 3, 4 and 5 present the results of these tests. For all tests we have used the same metrics previously identified as potential indicators of quality for each subset (Table 2).

Table 3 Results of DMCA for *Simulation* ∩ *Science and Technology* in the context of peer-reviews ratings thresholds

Classification algorithm	N	Classes in the model	Metrics used by the model[a]	Number of leaves				Classification precision				
				Number of rules	Size of the tree	MAE	K	Good (%)	Average (%)	Poor (%)	Not-good (%)	Overall (%)
J48	1	Good and not-good	2	3	5	0.31	0.38	33.33	–	–	98.43	76.29
	2	Good, average and poor	11	19	37	0.1	0.83	96.96	84.00	92.85	–	89.69
Simple cart	3	Good and not-good	2	3	5	0.30	0.53	57.57	–	–	92.18	80.41
	4	Good, average and poor	8	14	27	0.15	0.76	90.90	86.00	71.40	–	85.57
PART	5	Good and not-good	5	4	–	0.28	0.38	33.33	–	–	98.43	76.29
	6	Good, average and poor	8	11	–	0.16	0.74	97.00	72.00	92.9	–	83.51
Multilayer perceptron	7	Good and not-good	13	–	–	0.29	0.58	60.60	–	–	93.75	82.47
	8	Good, average and poor	13	–	–	0.26	0.53	60.60	92.00	42.90	–	74.23
Bayesian network	9	Good and not-good	3	–	–	0.30	0.37	84.84	–	–	57.81	67.01
	10	Good, average and poor	5	–	–	0.30	0.41	60.60	48.00	100	–	59.79

[a]All models were tested with 13 metrics.

Table 4 Results of DMCA for *Simulation* ∩ *Mathematics and Statistics* in the context of peer-reviews ratings thresholds

Classification algorithm	N	Classes in the model	Metrics used by the model[a]	Number of leaves				Classification precision				
				Number of rules	Size of the tree	MAE	K	Good	Average (%)	Poor	Not-good (%)	Overall (%)
J48	1	Good and not-good	2	4	7	0.36	0.44	58.1	–	–	84.60	74.70
	2	Good, average and poor	4	8	15	0.26	0.47	64.5	89.10	0	–	73.49
Simple cart	3	Good and not-good	1	2	3	0.4	0.37	48.4	–	–	86.50	72.29
	4	Good, average and poor	1	2	3	0.32	0.32	48.4	87.00	0	–	66.26
PART	5	Good and not-good	5	5	–	0.3	0.55	54.8	–	–	96.20	80.72
	6	Good, average and poor	5	–	–	0.23	0.55	77.4	87.00	0	–	77.11
Multilayer perceptron	7	Good and not-good	11	–	–	0.42	0.17	16.1	–	–	98.10	67.47
	8	Good, average and poor	11	–	–	0.34	0.13	16.1	97.80	0	–	60.24
Bayesian network	9	Good and not-good	0	–	–	0.47	0	0	–	–	100	62.65
	10	Good, average and poor	0	–	–	0.37	0	0	100	0	–	55.42

[a]All models were tested with 11 metrics.

Table 5 Results of DMCA for *Tutorial* ∩ *Science and Technology* in the context of peer-reviews ratings thresholds

Classification algorithm	N	Classes in the model	Metrics used by the model[a]	Number of leaves				Classification precision				
				Number of rules	Size of the tree	MAE	K	Good (%)	Average (%)	Poor (%)	Not-good (%)	Overall (%)
J48	1	Good and not-good	3	6	11	0.25	0.62	60.7	–	–	96.4	84.34
	2	Good, average and poor	2	4	7	0.37	0.21	0	97.2	47.4	–	53.01
Simple cart	3	Good and not-good	0	1	1	0.45	0	0	–	–	100	66.26
	4	Good, average and poor	5	10	19	0.24	0.64	82.1	83.3	57.9	–	77.11
PART	5	Good and not-good	4	6	–	0.24	0.66	67.9	–	–	94.5	85.54
	6	Good, average and poor	5	3	–	0.35	0.25	0	100	52.6	–	55.42
Multilayer perceptron	7	Good and not-good	13	–	–	0.40	0	0	–	–	100	66.26
	8	Good, average and poor	13	–	–	0.38	0.20	10.7	86.1	47.4	–	51.81
Bayesian network	9	Good and not-good	0	–	–	0.45	0	0	–	–	100	66.26
	10	Good, average and poor	0	–	–	0.43	0	0	100	0	–	43.37

[a]All models were tested with 13 metrics.

There are several possible criteria for evaluation the good prediction of classification models [16]. Here we selected a few of them to present the results of our analysis. In the tables, the column "metrics used by the model" presents the number of metrics that were included in the model generated by the given algorithm. The mean absolute error (MAE) measures the average deviation between the predicted classes and the true classes of the resources. The closer to 0 the MAE, the lower is the error of the prediction and the better the model. The K stands for "Kappa statistic" which is a coefficient that measures the overall agreement between the data observed and the data expected. This coefficient varies from −1 to 1, where 1 means total agreement, 0 means no agreement, and −1 means total disagreement. At last, the tables also present the overall precision of the model and the specific precisions for each one of the classes in the dataset. We adopted the MAE measure as the main reference of quality for the models, i.e., when we mention in this section that a given model is the best for a given subset, we mean that this model has presented the minimum MAE among all. In this first exploratory study the models were evaluated using the training dataset, i.e., the entire dataset was used for training and for evaluating.

As it can be seen in the tables, apparently there is no best classification algorithm that fits for all subsets for the generation of good models. The results vary significantly depending on the algorithm used, the subset from which the models were generated and the classes of quality included in the datasets.

Simulation ∩ Science and Technology

Among the three subsets, the models presented (in general) the best results for the *Simulation ∩ Science and Technology* subset. For this subset, the best model was a decision tree generated by a J48 algorithm (model number 2 of Table 3) which was able to correctly classify resources among *good*, *average* and *poor* with an overall precision of 89.69 %, and presented a Kappa coefficient of 0.83, and a MAE of just 0.1. The percentages of precision of this model for classifying resources in the specific categories of quality are considerably similar. *Good* resources are classified with 96.96 % of precision, while *average* and *poor* resources are classified with precisions of 84 and 92.85 % respectively. The second and third best models for this subset were also focused on classify resources among *good*, *average* and *poor*. The second best model was a decision tree generated by a Simple Cart algorithm with an overall precision of 85.57 % (model number 4 of Table 3) and the third best model was a set of if-then-rules generated by the PART algorithm with an overall precision of 83.51 % (model number 6 of Table 3). The main difference between these two models (in terms of precisions) is that the former presented the worst precision percentages for classifying *poor* resources (71.40 %), where the latter presented the worst precision percentages for classifying *average* resources (72 %). At last, the best results for classifying resources between *good* and *not-good* were achieved by the PART algorithm and by a Multilayer Perceptron Neural Network. The PART model achieved an overall precision of 76.29 a MAE of 0.28 and Kappa Statistic of 0.38. Moreover, it classified *not-good* resources with a precision of 98.43 %, and *good* resources with

a precision of only *33.33 %*. The Multilayer Perceptron presented an overall precision of 82.47 %, a MAE of 0.29 and a Kappa coefficient of 0.58. The drawback of these two models is the very low precision for classifying *good* resources.

Simulation ∩ Mathematics and Statistics

For the *Simulation ∩ Mathematics and Statistics* subset the best model was generated by the PART algorithm (model 5 of Table 4) for classifying resources between *good* and *not-good*. This model contains a set of 5 if-then-rules that uses 5 from the 11 metrics identified as possible indicators of quality. It achieved an overall precision of 80.72 %, a MAE of 0.30 and a Kappa coefficient equals to 0.55. Even though the overall results can be considered good, the model presents a serious limitation for the classification of *good* resources, with only 54.8 % of precision. The second best model for this subset is a decision tree generated by the J48 algorithm to classify resources between *good* and *not-good* (model 1 of the Table 4). Here the model achieved an overall precision of 74.70, a MAE of 0.36, and a Kappa coefficient of 0.44. The main problem with this model is the fact that it uses just 2 of the 11 possible indicators of quality. For this subset, all models for classifying resources among *good*, *average* and *poor* have completely failed on the classification of the *poor* category (presenting 0 % of precision). It is also possible to see that the precisions for classifying *good* and *average* resources in these models are very similar to the precisions for classifying *good* and *not-good* resources on the other models.

Tutorial ∩ Science and Technology

The best model for the subset *Tutorial ∩ Science and Technology* was generated by the PART algorithm to classify resources between *good* and *not-good* (model 5 of Table 5). The model presents an overall precision of 85.54 %, a MAE of 0.24 and a Kappa coefficient of 0.66. From the 13 metrics identified as quality indicators, the model has included only four in the six if-then-rules generated. Moreover, the model has a high precision for classifying *not-good* resources (94.5 %), but a low precision for classifying *good* resources (67.9 %). The second best model for this subset is a decision tree generated by a Simple Cart algorithm that classifies resources among *good*, *average* and *poor* (model 4 of Table 5). Here the model uses 5 from the 13 metrics identified as quality indicators; it has an overall precision of 77.11 %, a MAE of 0.24, and a Kappa coefficient of 0.64. The model is able to classify *good* resources with 82.1 % of precision, *average* resources with 83.3 % of precision, and *poor* resources with 57.9 % of precision. The third best model is a decision tree generated by a J48 algorithm (model 1 of Table 5). This model classifies resources between *good* and *not-good* with an overall precision of 84.34 %, a MAE of 0.25, and a Kappa coefficient of 0.62. The model uses only 3 from the 13 metrics identified as quality indicators. Moreover, similarly to the best model for this subset, this model also has a high precision for classifying *not-good* resources (96.4 %) and a low precision for classifying *good* resources (60.7 %).

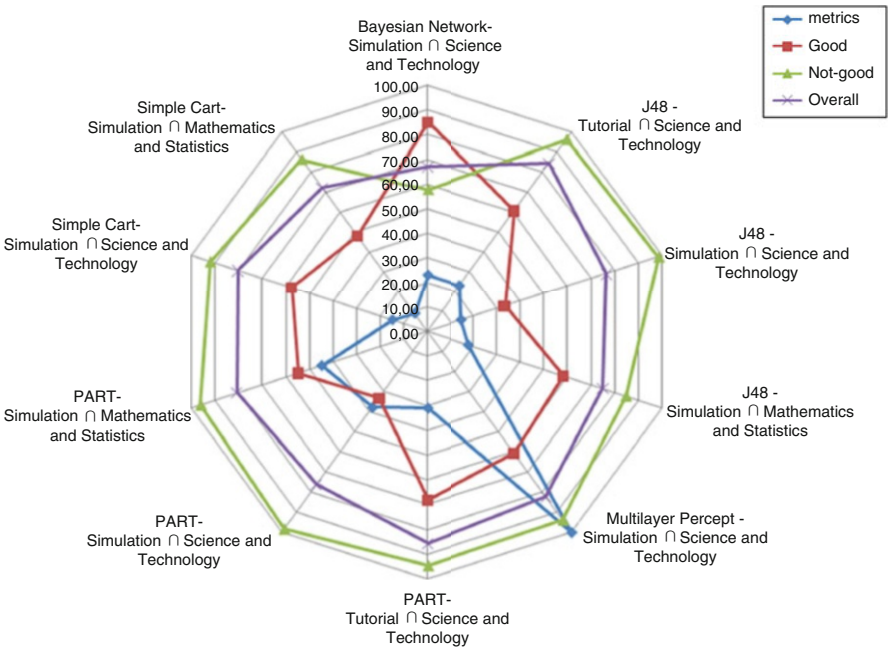

Fig. 2 Results of DMCA for *Tutorial* ∩ *Science and Technology* in the context of peer-reviews ratings thresholds

General Considerations at the light of the Results

The models normally exclude several of the metrics previously identified as indicators of quality. For instance, from the top ten best models for the classification of resources between Good and Not-Good, only one has used all metrics included in the dataset (a Multilayer Perceptron for the *Simulation* ∩ *Science and Technology* subset) (see Fig. 2). The rest of the models have used from just one to five metrics. It is also interesting to highlight that it was possible to generate models for all three subsets. Moreover, practically all models presented a higher precision for the classification of *not-good* resources than for *good* resources. Figure 2 presents this last observation more clearly. As it can be seen in the figure, from the ten best models, nine presented better precisions for classifying *not-good* resources and just one—a Bayesian Network for the *Simulation* ∩ *Science and Technology* subset—presented a higher precision for classifying *good* resources than *not-good* ones.

The best models generated for classifying resources among *good*, *average* and *poor* achieved lower MAEs and higher Kappa coefficients than the models for classifying resources between *good* and *not-good*. Moreover, as it can be seen in Fig. 3, the models here also tend to use more indicators of quality. The main problem found for this set of models is the fact that it was not possible to create good models for the subset of *Simulation* ∩ *Mathematics and Statistics* (all models presented 0.0 % of precision for

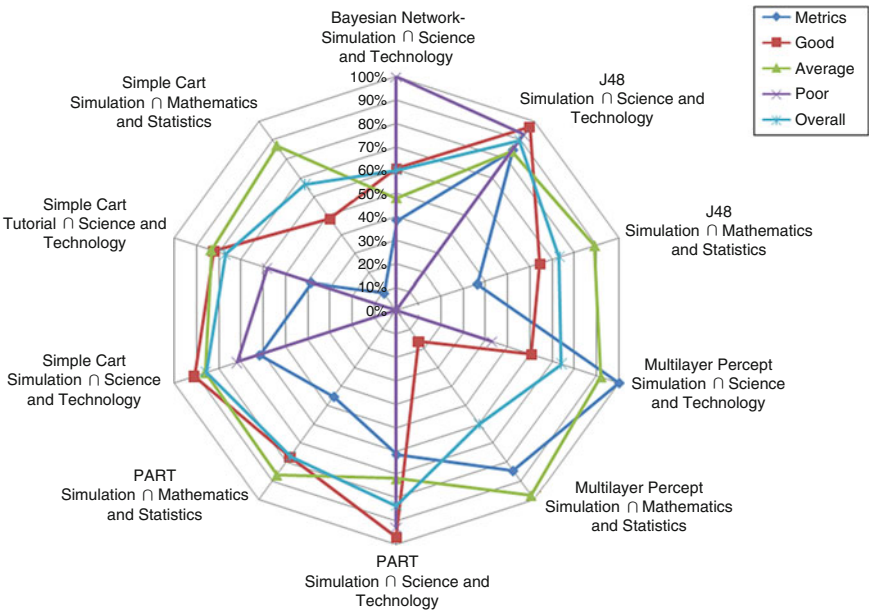

Fig. 3 Radar graph for the ten best models for classifying resources among *good*, *average* and *poor*

classifying *poor* resources). Another important thing to highlight is that the best three models presented more balanced precisions for the classification among the different classes. However, it is still possible to observe all kinds of models, i.e., those which classify more precisely *good* resources, those which classify more precisely *average* resources, and those which classify more precisely *poor* resources (see Fig. 3).

The results found here point out the possibility of generating models for automated quality assessment of learning resources inside repositories based on their intrinsic metrics. However, as the models are very heterogeneous (different MAEs, Kappa coefficients, number of metrics used, classification precisions), the decision of which one is the best will depend on the combination of several facts such as: the specific scenario to which the model is going to be applied, the specific subset (category of discipline versus material type) to which they are being generated for, and the classes of quality included in the dataset. Next section will describe another experiment towards automated evaluation and that was performed with a slightly different methodology and using a broader set of resources and subsets.

Second Experiment: Algorithmic Approach

For this second experiment we collected (in 2010) a total of 20,582 learning resources from MERLOT. From this amount, only 2,076 were peer-reviewed, and 5 of them did not have metadata regarding the category of discipline or the type of

Table 6 Frequency of materials for the subsets used in this study (intersection of category of discipline and material type)

Material type/discipline	Arts	Business	Education	Humanities
Collection		52	56	43
Reference material		83	40	51
Simulation	57	63	40	78
Tutorial		76	73	93
Material type/ discipline	Mathematics and statistics	Science & technology	Social sciences	
Collection	50	80		
Reference Material	68	102		
Simulation	40	150		
Tutorial	48	86		

material and were disregarded. Considering that many subsets are formed by very small amounts of resources, we restrained our experiment to just a few of them. Precisely, we worked with 21 subsets formed by the following types of material: *Collection*, *Reference Material*, *Simulation* and *Tutorial*, and that had 40 resources or more.[3] In total, we worked with information of 1,429 learning resources which represent 69 % of the total collected data. Table 6 presents the frequency of the materials for each subset used in this study.

As mentioned before, the methodology we followed for this experiment was slightly different from the one described in the previous section. Here we did not created statistical profiles of the learning resources, but used all collected metrics as input information for the generation and evaluation of models through the use of Artificial Neural Networks (ANNs).

This experiment was conducted with the Neural Network toolbox of Matlab. For each subset we randomly selected 70 % of the data for training, 15 % for testing and 15 % for validation, as suggested by Xu et al. [17]. We tested the Marquardt–Levenberg algorithm [18] using from 1 to 30 neurons in all tests. In order to obtain more statistically significant results (due to the small size of the data samples), each test was repeated 10 times and the average results were computed. Differently from the previous experiment, the models here were generated to classify resources between *good* and *not-good* (we did not tested models to classify resources among *good*, *average* and *poor*).

The choice of using ANNs rests on the fact that they are adaptive, distributed, and highly parallel systems which have been used in many knowledge areas and have proven to solve problems that require pattern recognition [19]. Moreover, ANNs are among the types of models that have also shown good precisions for some subsets in the previous experiment. At last, this experiment was initially

[3] The difficulties for training, validating and testing predictive models for subsets with less than 40 resources would be more severe.

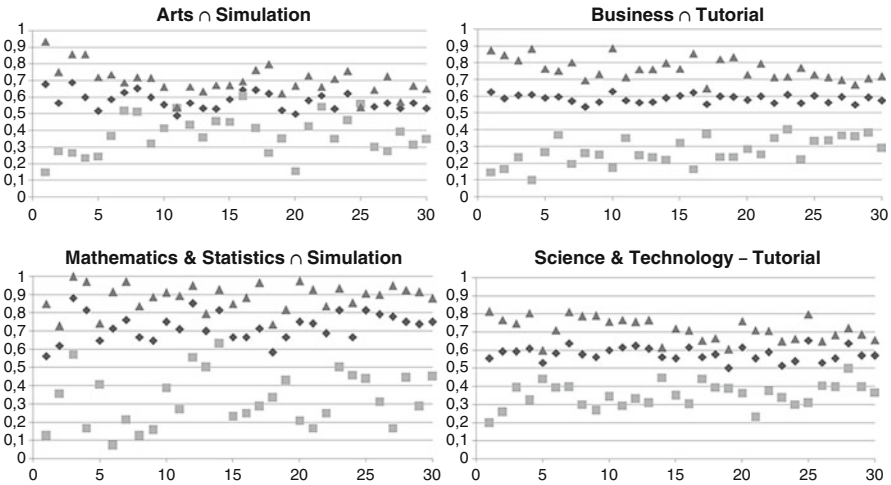

Fig. 4 Precisions of the some models versus number of neurons. Overall precision (*lozenges*), precision for the classification of *good* resources (*squares*) and *not-good* resources (*triangles*)

focused on populating the repository with hidden internal quality information that can be further used by ranking mechanisms [20], and for such a purpose we could use black-box models such as ANNs.

Results and Discussion

The models presented different results depending on the subset used for training. Most of the models tend to classify *not-good* resources better than *good* ones which can probably be a result of the uneven amount of resources of each class inside the datasets (normally formed by 2/3 of *not-good* and 1/3 of *good*). These tendencies can be observed in Fig. 4.[4]

The number of neurons used in the construction of the models has different influences depending on the subsets. A Spearman's rank correlation (r_s) analysis was carried out to evaluate whether there are associations between the number of neurons and the precisions achieved by the models. This test serves to the purpose of observing the pattern expressed by the models on predicting quality for the given subsets. For instance, assuming x as a predictive model for a given subset A, and y as a predictive model for a given subset B; if x has less neurons than y and both have the same precisions, the patterns expressed in A are simpler than the ones expressed in B. This means to say that it is easier to understand what is *good* (or *not-good*) in the subset A. Table 7 shows the results of such analysis.

[4] Just some models were presented in the figure.

Table 7 Tendencies of the precisions according to the number of neurons used for training (*good* | *not-good*)

Subset	Arts	Business	Education	Humanities	Math & statistics	Science & tech
Collection		– | –	↑ | ↓	– | –	– | –	– | –
Reference material		– | –	– | –	– | ↓	– | –	– | –
Simulation	– | ↓	↑ | –	– | ↓	– | –	– | –	↑ | ↓
Tutorial		↑ | ↓	↑ | ↓	↑ | –	– | –	– | ↓

In Table 7 (–) stands for no association between the number of neurons and the precision of the model for classifying a given class, (↑) stands for a positive association, and (↓) stands for a negative association. The analyses considered a 95 % level of significance. As it can be seen in the table, the number of neurons influences on the precisions for some classes of quality of some subsets. For instance, the number of neurons presents a positive association with the precisions for classifying *good* resources in the 6 (six) following subsets: *Business* ∩ *Simulation*, *Business* ∩ *Tutorial*, *Education* ∩ *Collection*, *Education* ∩ *Tutorial*, *Humanities* ∩ *Tutorial*, and *Science & Technology* ∩ *Simulation*. Moreover, the number of neurons presents a negative association with the precisions for classifying *not-good* resources in the 8 (eight) following subsets: *Arts* ∩ *Simulation*, *Business* ∩ *Tutorial*, *Education* ∩ *Collection*, *Education* ∩ *Simulation*, *Education* ∩ *Tutorial*, *Education* ∩ *Humanities*, *Science & Technology* ∩ *Simulation*, and *Science & Technology* ∩ *Tutorial*. Finally, there are no positive associations between the number of neurons and the precisions for classifying *not-good* resources; neither there are negative associations between the number of neurons and the precisions for classifying *good* resources.

In order to evaluate how to select the best models for quality assessment, it is necessary to understand the behavior 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 precisions 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. Most of the models have presented strong negative correlations between the precisions for classifying *good* and *not-good* resources. The results of both analyses suggest that the decision of selecting a model for predicting quality must take into account that, as the precision for classifying resources from one class increases, the precision for classifying resources of the other class decreases. Considering that, the question lies on establishing which would be the cutting point for acceptable precisions so that the models could be used for our purpose. In other words, it is necessary to establish the minimum precisions (cutting point) that the models must present for classifying both classes (*good* and *not-good*) so that they can be used for generating hidden quality information for the repository.

For the present study, we are considering that the models must present precisions higher than 50 % for the correct classification of *good* and *not-good* resources (simultaneously) in order to be considered as useful. It is known that the decision of selecting the minimum precisions for considering a model as efficient or not will depend on

Table 8 Two best models for each subset (ordered by the precisions for classifying *good* resources)

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

the specific scenario/problem for which the models are being developed for. Here we are considering that precisions higher than 50 % are better than the merely random.

Table 8 presents the top-2 models for each subset considering their overall precisions, and their precisions for classifying *good* and *not-good* resources (ordered by the precision for classifying *good* resources).

In Table 8, N stands for the number of neurons in the model, OP stands for the overall precision, G for the precision for classifying good resources and NG for the precision for classifying not-good resources. As it can be seen in the table, and considering the established minimum cutting-point, it was possible to generate models for almost all subsets. From the 42 models presented in the table, only 10 did not reach the minimum precisions (white in the table). Moreover, 22 of them presented precisions between 50 and 59.90 % (gray hashed in the table), and nine presented both precisions higher than 60 % (black hashed in the table). We have also found 1 (one) model with precisions higher than 70 % (for *Humanities* ∩ *Simulation*). The only three subsets where the models did not reach the minimum precisions were: *Business* ∩ *Tutorial*, *Education* ∩ *Collection* and *Education* ∩ *Tutorial*. On the other hand, the best results were found for: *Humanities* ∩ *Simulation, Mathematics* ∩ *Tutorial, Humanities* ∩ *Collection, Business* ∩ *Simulation*, *Arts* ∩ *Simulation* and *Business* ∩ *Collection.* One of the possible reasons why it was not feasible to generate good models for all subsets may rest on the fact that the real features associated to quality on those given subsets might not have been collected by the crawler.

In order to select the most suitable model one should take into consideration that the model's output is going to be used as information during the ranking process, and to evaluate the advantages and drawbacks of a lower precision for classifying *good* resources in contraposition to a lower precision for classifying *not-good* resources. The less damaging situation seems to occur when the model classifies as *not-good* a *good* material. In this case, *good* materials would just remain hidden in the repository, i.e., in bad ranked positions (a similar situation to the one of not using the models). On the other hand, if the model classifies as *good* a resource that is *not-good*, it is most likely that this resource will be put at a higher rank position, thus increasing its chances of being accessed by the users. This would mislead the user towards the selection of a "not-so-good" quality resource, and it could put in discredit the ranking mechanism.

Conclusions and Outlook

It is known that LORs normally use evaluative information to rank resources during the process of search and retrieval. However, the amount of resources inside LORs increases more rapidly than the number of contributions given by the community of users and experts. Because of that, many LOs that do not have any quality evaluation receive bad rank positions even if they are of high-quality, thus remaining unused (or unseen) inside the repository until someone decides to evaluate it.

The present chapter presented two experiments that used intrinsic features of the resources in order to generate models for their automated quality assessment. For that, we collected information from MERLOT and used the ratings associated to the resources as baseline for the creation of classes of quality.

In the first experiment we tested the generation of automated models through the creation of statistical profiles and the further use of data mining classification algorithms for three distinct subsets of MERLOT materials. On these studies we were able to generate models with good overall precision rates (up to 89 %) but we highlighted that the feasibility of the models will depend on the specific method used to generate them, the specifics subsets to which they are being generated for, and the classes of quality included in the dataset. Moreover, the models were generated by using considerably small datasets (around 90 resources each), and were evaluated using the training dataset, i.e., the entire dataset was used for training and for evaluating. Such kind of evaluation is always too optimistic and is susceptible to over fitting (i.e. the model just memorizes the data and can fail to predict well in the future).

In the second experiment we used all collected intrinsic features as input information for the generation of models represented by Artificial Neural Networks. We also changed the method for the evaluation of the models in order to better deal with the small amount of resources in the samples and to avoid over fitting. Among other good results, one can mention the model for *Humanities* ∩ *Simulation* that is able to classify *good* resources with 75 % of precision and *not-good* resources with 79 %; and the model developed for *Mathematics* ∩ *Tutorial* with 79 % of precision

for classifying *good* resources and 64 % for classifying *not-good* ones. As the models would be used inside repository and the classifications would serve just as input information for searching mechanisms, it is not necessarily required that the models provide explanations about their reasoning. Models constituted of neural networks (as the one tested in the present study) can perfectly be used in such a scenario.

The models developed here could be used to provide internal quality information for those LOs still not evaluated, thus helping the repository in the stage of offering resources. Resources recently added to the repository would be highly benefited by such models since that they hardly receive any assessment just after their inclusion. Once the resource finally receives a formal evaluation from the community of the repository, the initial implicit quality information provided by the model could be disregarded. Moreover, this "real" rating could be used as feedback information so that the efficiency of the models could be analyzed, i.e. to evaluate whether or not the users agree with the models decisions.

Future work will try to include more metrics still not implemented, such as, for instance, the number of colors and different font styles, the existence of adds, the number of redundant and broken links, and some readability measures (e.g. Gunning Fog index and Flesch-Kincaid grade level). We would also like to repeat the experiments, but now using the same method to train and evaluate the models so that we can compare the results of these two approaches. Besides, as pointed out by Cechinel and Sánchez-Alonso [21], both communities of evaluators in MERLOT (users and peer-reviewers) are communicating different views regarding the quality of the learning objects refereed in the repository. The models tested here are related to the perspective of quality given by peer-reviewers. Future work will test models created with the ratings given by the community of users and will compare their performances with the present study. Moreover, as the present work is context sensitive, it is important to evaluate whether this approach can be extended to other repositories. As not all repositories adopt the same kind of quality assurance that MERLOT does, alternative quality measures for contrasting classes between *good* and *not-good* resources must be found. Another interesting possible direction is to classify learning resources according to their granularity, and use this information as input for the generation of the models. At last, we could use the values calculated by the models for all the resources and compare the ranking of MERLOT with the ranking performed through the use of these "artificial" quality information.

It is important to mention that the present approaches do not intend to replace traditional evaluation methods, but complement them providing a useful and inexpensive quality assessment that can be used by the repositories before more time and effort consuming evaluation is performed.

Acknowledgments The work presented here has been partially funded by the European Commission through the project IGUAL (www.igualproject.org)—Innovation for Equality in Latin American University (code DCIALA/19.09.01/10/21526/245-315/ALFAIII (2010)123) of the ALFA III Programme, by Spanish Ministry of Science and Innovation through project MAVSEL: Mining, data analysis and visualization based in social aspects of e-learning (code TIN2010-21715-C02-01) 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

1. 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. Idea Group, Hershey, PA, pp 87–107
2. Ochoa X, Duval E (2009) Quantitative analysis of learning object repositories. IEEE Trans Learn Technol 2(3):226–238
3. Ochoa X, Duval E (2008) Relevance ranking metrics for learning objects. IEEE Trans Learn Technol 1(1):34–48. doi:10.1109/TLT.2008.1, http://dx.doi.org/
4. Sanz-Rodriguez J, Dodero J, Sánchez-Alonso S (2010) Ranking learning objects through integration of different quality indicators. IEEE Trans Learn Technol 3(4):358–363. doi:10.1109/TLT.2010.23
5. Cechinel C, Sánchez-Alonso S, Sicilia M-Á (2009) Empirical analysis of errors on human-generated learning objects metadata. In: Sartori F, Sicilia MÁ, Manouselis N (eds) Metadata and semantic research, vol 46, Communications in computer and information science. Springer, Berlin, pp 60–70. doi:10.1007/978-3-642-04590-5_6
6. Cechinel C, Sánchez-Alonso S, García-Barriocanal E (2011) Statistical profiles of highly-rated learning objects. Comput Educ 57(1):1255–1269. doi:10.1016/j.compedu.2011.01.012
7. Cechinel C, Silva Camargo S, Sánchez-Alonso S, Sicilia M-Á (2012) On the search for intrinsic quality metrics of learning objects. In: Dodero J, Palomo-Duarte M, Karampiperis P (eds) Metadata and semantics research, Communications in computer and information science. Springer, Berlin, pp 49–60. doi:10.1007/978-3-642-35233-1_5
8. Meyer M, Hannappel A, Rensing C, Steinmetz R (2007) Automatic classification of didactic functions of e-learning resources. Paper presented at the Proceedings of the 15th international conference on multimedia, Augsburg, Germany
9. Mendes E, Hall W, Harrison R (1998) Applying metrics to the evaluation of educational hypermedia applications. J Univers Comput Sci 4(4):382–403. doi:10.3217/jucs-004-04-0382
10. Blumenstock JE (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
11. Stvilia B, Twidale MB, Smith LC, Gasser L (2005) Assessing information quality of a community-based encyclopedia. In: Proceedings of the international conference on information quality – ICIQ 2005, pp 442-454. Doi:citeulike-article-id:1833325
12. Ivory MY, Hearst MA (2002) Statistical profiles of highly-rated web sites. Changing our world, changing ourselves. Paper presented at the proceedings of the SIGCHI conference on Human factors in computing systems, Minneapolis, MA, 2002
13. Nesbit JC, Belfer K, Leacock T (2003) Learning object review instrument (LORI). E-learning research and assessment network. http://www.elera.net/eLera/Home/Articles/LORI%20manual
14. García-Barriocanal E, Sicilia M-Á (2009) Preliminary explorations on the statistical profiles of highly-rated learning objects. In: Sartori F, Sicilia MÁ, Manouselis N (eds) Metadata and semantic research, vol 46, Communications in computer and information science. Springer, Berlin, pp 108–117. doi:10.1007/978-3-642-04590-5_10
15. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explor Newsl 11(1):10–18. doi:10.1145/1656274.1656278
16. Cichosz P (2011) Assessing the quality of classification models: performance measures and evaluation procedures. Cent Eur J Eng 1(2):132–158. doi:10.2478/s13531-011-0022-9
17. Xu L, Hoos HH, 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
18. Hagan MT, Menhaj MB (1994) Training feedforward networks with the Marquardt algorithm. IEEE Trans Neural Netw 5(6):989–993. doi:10.1109/72.329697
19. Bishop CM (2006) Pattern recognition and machine learning, Information Science and Statistics. Springer, New York

20. Cechinel C, Camargo SdS, Ochoa X, Sánchez-Alonso S, Sicilia M-Á (2012a) Populating learning object repositories with hidden internal quality information. In: Manouselis N, Drachsler H, Verbert K, Santos OC (eds) Recommender systems in technology enhanced learning, CEUR workshop proceedings, Saarbrücken, pp 11–22
21. Cechinel C, Sánchez-Alonso S (2011) Analyzing associations between the different ratings dimensions of the MERLOT repository. Interdisciplinary Journal of E-Learning and Learning Objects 7:1–9

http://www.springer.com/978-1-4939-0529-4

Recommender Systems for Technology Enhanced Learning
Research Trends and Applications
Manouselis, N.; Drachsler, H.; Verbert, K.; Santos, O.C. (Eds.)
2014, XIV, 306 p. 67 illus., Hardcover
ISBN: 978-1-4939-0529-4