Populating Learning Object Repositories with Hidden Internal Quality Information

Cristian Cechinel[1], Sandro da Silva Camargo[1], Xavier Ochoa[2], Salvador Sánchez-Alonso[3], Miguel-Ángel Sicilia[3]

[1] Computer Engineering Course
Federal University of Pampa, Caixa Postal 07
96400-970, Bagé (RS), Brazil
contato@cristiancechinel.pro.br, camargo.sandro@gmail.com

[2] Escuela Superior Politécnica del Litoral, Campus Gustavo Galindo, Km. 30. Vía Perimetral, Guayaquil, Ecuador, xavier@cti.espol.edu.ec

[3] Information Engineering Research Unit
Computer Science Dept., University of Alcalá
Ctra. Barcelona km. 33.6 – 28871 Alcalá de Henares (Madrid), Spain
salvador.sanchez, msicilia@uah.es

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 an experiment for automatically generate quality information about learning resources inside repositories through the use of Artificial Neural Networks models. We were able to generate models for classifying resources between *good* and *not-good* with accuracies that vary from 50% to 80% depending on the given subset. The preliminary results found here point out the feasibility of such approach and can be used as a starting point for the pursuit of automatically generation of internal quality information about resources inside repositories.

Keywords: Ranking mechanisms; ratings; learning objects; learning object repositories; MERLOT; Artificial Neural Networks

1 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 individual needs. Inside LORs, this kind of evaluative metadata (Vuorikari, Manouselis, & Duval, 2008) 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 (Ochoa & Duval, 2009) and it became 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 one month (from May 21th to June 21th), the amount of new resources catalogued in the repository was 9 times more than the amount of new ratings given by experts (peer-reviewers), 6 times more than the amount of new comments (and users ratings) and 3 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 disadvantage position to compete for a good place 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 (2008) developed a set of metrics for ranking the results of learning objects search 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. The 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, Dodero, and Sánchez-Alonso (2010) 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 measure quality (or calculate relevance) according to specific dimensions depend 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 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, Sánchez-Alonso, and García-Barriocanal (2011) who developed highly-rated profiles of learning objects available in the MERLOT repository, 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 accuracy, and between good and poor with 91.49% of accuracy. Among other things, the authors have 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 the MERLOT repository, 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. On an another round of experiments, Cechinel (2012) also tested the creation 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 the author were able to generate models with good overall precision rates (up to 89%) but the author 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.

The present work expands the previous works developed by Cechinel (2012) and Cechinel et al. (2011) by generating and evaluating models for automated quality assessment of learning objects stored on MERLOT focusing on populating the repository with hidden internal quality information that can be further used by ranking mechanisms. On the previous works, the authors explored the creation of statistical profiles of highly-rated learning objects by contrasting information from the *good* and *not-good* resources and then used these profiles to generate models for quality assessment. In the present work we are testing a slightly different and more algorithmic approach, i.e., the models here are being generated exclusively through the use of data mining algorithms. Moreover, we are also working with a larger collection of resources and a considerably higher number of MERLOT subsets. The rest of this paper is structured as follows. Section 2 presents existing research focused on identifying intrinsic quality features of resources. Section 3 describes the methodology followed for the study and section 4 discusses the results found. Finally, conclusions and outlook are provided in Section 5.

2 Background

From our knowledge, besides the recent work of Cechinel et al. (2011), 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 (Meyer, Hannappel, Rensing, & Steinmetz, 2007), 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, Hall, and Harrison (1998) 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 (2008) has found the length of an article (measured in words) as a predictor of quality in Wikipedia. Moreover, Stvilia, Twidale, Smith & Gasser (2005) 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 (2002) 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 (Nesbit, Belfer, & Leacock, 2003) 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.

3 Methodology

The main objective of this research was to obtain models that could automatically identify *good* and *not-good* learning objects inside repositories based on the intrinsic features of the resources. The methodology that we followed was the development of models though the use of data mining algorithms over information of learning objects catalogued on MERLOT repository. For that, a database was collected from the repository and qualitative classes of quality of *good* and *not-good* were generated considering the terciles of the ratings of the resources. These classes of quality were then used as the reference output for the generation of the models.

3.1 Data Collection

A database was collected from MERLOT through the use of a crawlerthat systematically traversed the pages and collected information related to 35 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. (2011) and some of them have also been mentioned in other works which tackled the problem of assessing quality of resources (previously presented in section 2).

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

As resources in MERLOT vary considerably in size, a limit of 2 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. For instance, an object composed by 3 pages and containing a total of 30 images, will have a total number of images of 30, and an average number of images equals to 10 (30/3). Information of a total of 20,582 learning resources was collected. 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 material and were disregarded. Considering that many subsets are formed by very small amount 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[2]. In total, we worked with information of 1,429 learning resources which represent 69% of the total collected data. Table 2 presents the frequency of the materials for each subset used in this study.

[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 sample used in this study, the crawler kept running uninterruptedly for 4 full months.

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

Table 2: 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	

3.2 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[3]. 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.*

3.3 Mining models for automated quality classification of learning objects

The classes of quality were used as the output reference for generating and testing models for automated quality assessment of the resources through the use of Artificial Neural Networks (ANNs). 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 (Bishop, 2006). Moreover, ANNs are among the types of models that have also shown good accuracies on the previous works mentioned before. Finally, we have initially tested other approaches (with rules and trees) and they presented maximum accuracies around 60%. As ANNs presented the best preliminary results we selected this approach for the present study.

The experiments were 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, Hoos, and Leyton-Brown (2007). We tested the Marquardt –Levenberg algorithm (Hagan & Menhaj, 1994) 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. The models were generated to classify resources between *good* and *not-good.*

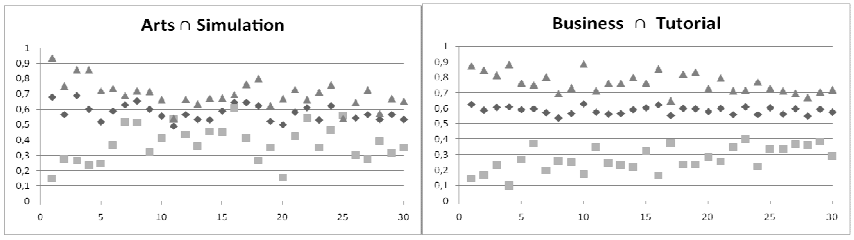
[3] The terciles of each subset were omitted from the paper due to a lack of space

4 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 figure 2[4].

The number of neurons used on 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 accuracies 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 accuracies, 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 3 shows the results of such analysis.

In Table 3 (-) stands for no association between the number of neurons and the accuracy 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 accuracies for some classes of quality of some subsets. For instance, the number of neurons presents a positive association with the accuracies 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 accuracies 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 accuracies for classifying *not-good* resources; neither there are negative associations between the number of neurons and the accuracies for classifying *good* resources.



[4] Just some models were presented in the figure due to a lack of space

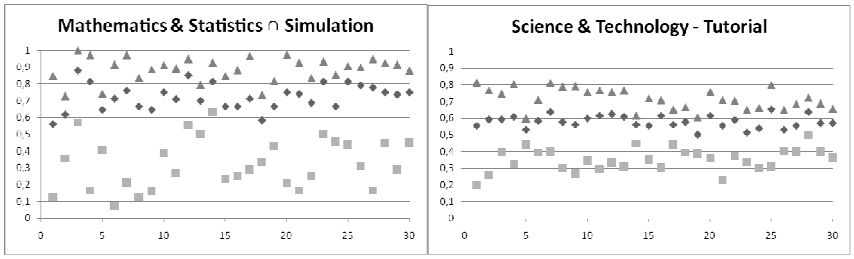


Fig.2. Accuracies of the some models versus number of neurons. Overall accuracies (lozenges), accuracies for the classification of *good* resources (squares) and *not-good* resources (triangles)

Table 3: Tendencies of the accuracies 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 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 on 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. Most of the models have presented strong negative correlations between the accuracies 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 accuracy for classifying resources from one class increases, the accuracy for classifying resources of the other class decreases. Considering that, the question lies on establishing which would be the cutting point for acceptable accuracies so that the models could be used for our purpose. In other words, it is necessary to establish the minimum accuracies (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 accuracies 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 accuracies for considering a model as efficient or not will depend on the specific scenario/problem for which the models are being developed for. Here we are considering that accuracies higher than 50% are better than the merely random.

Table 4 presents the top-2 models for each subset considering their overall accuracies, and their accuracies for classifying *good* and *not-good* resources (ordered by the accuracy for classifying *good* resources).

Table 4: Two best models for each subset (ordered by the accuracies for classifying *good* resources)

Subset	*N*	*OA*	*G*	*NG*	*Subset*	*N*	*OA*	*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
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
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
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
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
Humanities ∩ Reference Mat.	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
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
Math. ∩ Reference Mat.	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
Mathematics ∩ 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
Science & Tech. ∩ Reference Mat.	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
Science & Tech. ∩ Tutorial	28	0,64	0,50	0,72					
	14	0,56	0,45	0,61					

In table 4, N stands for the number of neurons in the model, OA stands for the overall accuracy, G for the accuracy for classifying good resources and NG for the accuracy 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 accuracies (white in the table). Moreover, 22 of them presented accuracies between 50% and 59.90% (gray hashed in the table), and 9 presented both accuracies higher than 60% (black hashed in the table). We have also found 1 (one) model with accuracies higher than 70% (for *Humanities ∩ Simulation*). The only three subsets to which the models did not reach the minimum accuracies 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 accuracy for classifying *good* resources in contraposition to a lower accuracy for classifying *not-good* resources. The less damaging situation seems to occur when the model classify 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.

5 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 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. 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.

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). Besides, as pointed out by Cechinel and Sánchez-Alonso (2011), 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 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 one of the metrics to be evaluated during the creation of the highly-rated profiles. 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 approach does 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 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/ALFAHI (2010)123) of the ALFA III Programme, and 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).

References

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*: Springer.

Blumenstock, Joshua 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.

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

Cechinel, Cristian, & Sánchez-Alonso, Salvador. (2011). Analyzing Associations between the Different Ratings Dimensions of the MERLOT Repository. *Interdisciplinary Journal of E-Learning and Learning Objects 7*, 1-9.

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

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

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.

Mendes, Emilia, Hall, Wendy, & Harrison, Rachel. (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

Meyer, Marek, Hannappel, Alexander, Rensing, Christoph, & Steinmetz, Ralf. (2007). *Automatic classification of didactic functions of e-learning resources*. Paper presented at the Proceedings of the 15th international conference on Multimedia, Augsburg, Germany.

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

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

Sanz-Rodriguez, Javier, Dodero, Juan, & Sánchez-Alonso, Salvador. (2010). Ranking Learning Objects through Integration of Different Quality Indicators. *IEEE Transactions on Learning Technologies, 3*(4), 358 - 363. doi: 10.1109/TLT.2010.23

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, Riina, Manouselis, Nikos , & Duval, Erik. (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, Lin, Hoos, Holger H., & Leyton-Brown, Kevin. (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.