View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Archivio della ricerca- Università di Roma La Sapienza

Mining Prerequisite Relationships Among Learning Objects

Carlo De Medio[1,2], Fabio Gasparetti[1], Carla Limongelli[1], Filippo Sciarrone[1], and Marco Temperini[2]

[1] Department of Engineering, Roma Tre University
Via della Vasca Navale 79 - 00146 Rome, Italy
carlo.demedio@uniroma3.it {limongel,gaspare,sciarro}@ing.uniroma3.it
[2] Department of Computer, Control and Management Engineering
Sapienza University, Via Ariosto, 25 - 00184 Roma, Italy
marte@dis.uniroma1.it

Abstract. The process of carefully choosing and sequencing a set of Learning Objects (LOs) to build a course may reveal to be quite a challenging task. In this work we focus on an aspect of such challenge, related to the verification and respect of the relationships of pedagogical dependence that holds between two LOs added to a course (meaning that if a given LO has another one as "pre-requisite", then any sequencing of the LOs in the course will need to have the latter LO taken by the learners before of the former). An innovative Machine learning-based approach for the identification of these kinds of relationships is proposed.

Keywords: learning objects, sequencing, prerequisite relationships

1 Introduction

In the case of online courses, a Learning Object (LO) can be seen as a digital object that is used for achieving a desired learning outcome or educational goal. With the ever-increasing use of learning management systems (LMS), repositories of LOs to be considered in specialized training are getting popular and heterogeneous w.r.t. covered disciplines. They encourage the instructors to adopt (and adapt) such LOs while building their education courses. Popular examples are Connexion[3], Ariadne[4] and Merlot[5]. Autonomous crawling techniques can also help building these repositories by sifting through hypertext resources on the web [1, 2].

Several factors inhibit a more widespread use of such paradigm of course development. Often LOs follow poorly, or not at all, the expected standardized

[3] Connexions is a Learning Object Repository, available at http://www.cnx.org (Accessed Jan- uary 27 2016)
[4] Ariadne Foundation, available at http://ariadne-eu.org (Accessed January 27 2016)
[5] Merlot is a Learning Object Repository, available at http://www.merlot.org (Accessed January 27 2016)

meta-tag or classification scheme, and badly needed references to other LOs are missing. This could negatively impair the vision of LOs as resources that, once created, can be quickly retrieved and used several times in different contexts, compensating the high cost of production.

Promising research activities are studying ontologies and Semantic Web technologies, allowing to address these issues, and capable to support the development of next generation LO repositories [3]; yet, creating education ontologies remains a time-consuming and error-prone task.

On the other side of the same coin, building an e-learning course by a sequence of LOs, i.e. by selecting didactic resources and designing their organization in the course, is a multi-level, multi-faceted and iterative process, in which different skills and knowledge are required. In this kind of task, recommender and filtering tools can be of substantial help [4–7].

In our approach the sequencing of LOs in the course can still be managed by the instructors, basing on their taste and preferences, yet they can be also helped by a set of suggestions, related to the pre-requisite relationships holding among the LOs selected for the course. Such relationships can be automatically computed and provide the instructor with significant help and guidance. We show a light-weight formalization of the LO, and how it can be represented by a set of WikiPedia articles; then we show how such set of topics can help deciding on the dependence relationship holding between two LOs. In this endeavor we exploit the classification in categories available for the WikiPedia articles, and obtain interesting results for our framework, in terms of precision and recall of the dependence relationships.

2 Related works

Wikipedia offers a quantity of high quality content resources in terms of presentation [8]. The openness, easy availability, and freshness of data make Wikipedia of interest in a variety of research activities, such as natural language processing and translation tools. Links, categories and information in templates provide structured content, which can be retrieved from raw XML dumps or Application Program Interface calls.

While some attempts aim at incorporating selected Wikipedia content into the curriculum as a collaborative environment [9] or for categorizing learning resources [10], to our knowledge our approach is novel w.r.t. inferring dependency relationships between LOs.

An interesting case-based reasoning approach, following a self-directed learning paradigm in assisting users to build sequences of elements out of user-defined libraries, is proposed in [11].

An evaluation of the hypotheses that motivated this research has been previously discussed in the following works: [12–14].

3 Mining Prerequisites

The current proposal consists in a traditional Machine Learning (ML) approach [15] applied to a dataset of LOs by performing a comparative analysis of several features of the LOs. The dataset is composed by LOs coming from five web-based courses we managed, on a wide variety of subject matters.

The presented approach is implemented in a software system that supports the following process. Firstly the set of LOs is textually analyzed, and each LO is associated to a Wikipedia page (*topic*): the set of topics is considered representative of the set of LOs. Then, the fact that a LO is represented by a topic allows to quantify the values of a set of features of the LO, by computing them on the associated topic.

We define the features according with peculiar aspects of the representative topics such as content length, generality, or specialization. Namely, given two learning objects LO_i and LO_j, we have: (1) the two average lengths of the text of the Wikipedia topics associated to the pair defined in terms of words obtained by a text tokenization process, (2) the number of links in the first section of the Wikipedia topics, (3) the average number of links in the topics associated to the LOs, (4) the number of distinct nouns in the LOs extracted by a part-of-speech tagger, (5) the intersection of the two sets of nouns extracted from the two LOs, (6) similar to the features #1 but limited to the first section of Wikipedia and (7) the intersection between the set of nouns used in links to other topics in the topics associated to LO_i, and the nouns extracted from LO_j.

So then, the topics are analyzed and the related LOs features computed. Finally the dependency relation between two LOs is inferred taking their features under consideration: this computation is obtained by feeding the features into a ML-based classifier.

4 Empirical Evaluation

In data mining, a decision tree is a predictive model that can be used to represent both classifiers and regression models. J48 is the implementation of C4.5 algorithm [16] developed by J. Ross Quinlan. C4.5 algorithm produces decision tree classification for a given dataset by recursive division of the data and the tree is grown using Depth-first strategy. Pruning methods have been introduced to reduce the complexity of tree structure without decreasing the accuracy of classification. Subtree raising is the followed pruning support procedure, that is, moving nodes upwards toward the root of tree and also replacing other nodes on the same way [17].

JRip is the propositional rule learner based on the Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [18]. Starting with the less prevalent classes, the algorithm iteratively grows and prunes rules until there are no positive examples left. It tries every potential value of each attribute and selects the condition with highest information gain. The minimum description length is considered as stopping criterion when new conditions are sequentially added to a rule.

These two ML algorithms have been considered for the the classification task, where the following measures can be defined:

- *tp*: the number of identified dependencies that are also expected in the test set;
- *fp*: the number of dependencies returned by the classifier but missing in the test set;
- *fn*: the number of expected dependencies that the classifier misses to identify.

and, consequently, the performances can be evaluated with the standard measures of Precision (**Pr**) and Recall (**Re**).

$$Pr = \frac{tp}{tp + fp} \qquad\qquad Re = \frac{tp}{tp + fn}$$

that is, the precision and the recall.

Five course materials with various levels of difficulty, conveying different random topics, e.g., scientific, archaeological, cinematography and art; have been considered for the evaluation. A domain expert manually identified the expected dependencies among LOs.

The average precision (**Pr**) reaches 0.828 and 0.736, for J48 and JRip, respectively. The recall (**Re**) values range from 0.811 (J48) and 0.756 (JRip). Each approach is validated following a 10-fold cross-validation. The outcomes prove that the hypothesis of a classifier trained on features extracted from two LOs has the chance to correctly identifying prerequisites among them.

5 Conclusions

We have presented and evaluated a Machine learning-based approach for mining prerequisite relations between learning objects. It can be used in a more comprehensive approach for helping teachers in searching relevant content and assisting them during the course development.

In our future work, we plan to continue evaluating the precision of the proposed approach in different domains of interest. In some circumstances (e.g., Mathematics and Statistics courses), the semantic annotation does not successfully associate relevant topics to the learning objects. Alternative approaches must be considered in order to overcome this issue and categorize the features exatracted from the LOs [19]. Preferences of teachers manifested through the course development can also be studied and combined, for example by monitoring the browsing behaviour on learning objects represented by hypertext resources [12].

References

1. Gasparetti, F., Micarelli, A.: Adaptive Web Search Based on a Colony of Cooperative Distributed Agents. In: Cooperative Information Agents VII: 7th International

Workshop, CIA 2003, Helsinki, Finland, August 27-29, 2003. Proceedings. Springer Berlin Heidelberg, Berlin, Heidelberg (2003) 168–183
2. Micarelli, A., Gasparetti, F.: Adaptive focused crawling. In Brusilovsky, P., Kobsa, A., Nejdl, W., eds.: The Adaptive Web. Volume 4321 of Lecture Notes in Computer Science. Springer-Verlag, Berlin, Heidelberg (2007) 231–262
3. Raju, P., Ahmed, V.: Enabling technologies for developing next-generation learning object repository for construction. Automation in Construction **22** (2012) 247 – 257 Planning Future Cities-Selected papers from the 2010 eCAADe Conference.
4. Limongelli, C., Sciarrone, F., Starace, P., Temperini, M.: An ontology-driven olap system to help teachers in the analysis of web learning object repositories. Information Systems Management **27**(3) (2010) 198–206
5. Limongelli, C., Lombardi, M., Marani, A., Sciarrone, F., Temperini, M.: A recommendation module to help teachers build courses through the moodle learning management system. New Review of Hypermedia and Multimedia (2015) Published online - Article in Press.
6. Limongelli, C., Sciarrone, F., Temperini, M.: A social network-based teacher model to support course construction. Computers in Human Behavior (2015) Article in Press.
7. Revilla Muñoz, O., Alpiste Penalba, F., Fernández Sánchez, J.: The skills, competences, and attitude toward information and communications technology recommender system: an online support program for teachers with personalized recommendations. New Review of Hypermedia and Multimedia (2015) Published online - Article in Press.
8. Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F.., Lanamki, A.: the sum of all human knowledge: A systematic review of scholarly research on the content of wikipedia. Journal of the Association for Information Science and Technology **66**(2) (2015) 219–245
9. Forte, A., Bruckman, A.: From wikipedia to the classroom: Exploring online publication and learning. In: Proceedings of the 7th International Conference on Learning Sciences. ICLS '06, International Society of the Learning Sciences (2006) 182–188
10. Meyer, M., Rensing, C., Steinmetz, R.: Categorizing Learning Objects Based On Wikipedia as Substitute Corpus. CEUR Workshop Proceedings. (September 2007)
11. Gasparetti, F., Micarelli, A., Sciarrone, F.: A web-based training system for business letter writing. Knowledge-Based Systems **22**(4) (May 2009) 287–291
12. Gasparetti, F., Micarelli, A., Sansonetti, G.: Exploiting web browsing activities for user needs identification. In: Computational Science and Computational Intelligence (CSCI), 2014 International Conference on. Volume 2. (March 2014) 86–89
13. Gasparetti, F., Limongelli, C., Sciarrone, F.: A content-based approach for supporting teachers in discovering dependency relationships between instructional units in distance learning environments. In Stephanidis, C., ed.: HCI International 2015 - Posters' Extended Abstracts, Los Angeles, CA, USA, August 2-7, 2015. Volume 529., Springer (2015) 241–246
14. Medio, C.D., Gasparetti, F., Limongelli, C., Sciarrone, F., Temperini, M.: Automatic extraction of prerequisites among learning objects using wikipedia-based content analysis. In: Proceedings of the 13th International Conference on Intelligent Tutoring Systems. ITS '16, Springer-Verlag (2016)
15. Mitchell, T.M.: Machine Learning. 1 edn. McGraw-Hill, Inc., New York, NY, USA (1997)

16. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2005)
17. Zhao, Y., Zhang, Y.: Comparison of decision tree methods for finding active objects. Advances in Space Research **41**(12) (2008) 1955 – 1959
18. Leon, F., Aignatoaiei, B., Zaharia, M.: Performance analysis of algorithms for protein structure classification. In: Database and Expert Systems Application, 2009. DEXA '09. 20th International Workshop on. (Aug 2009) 203–207
19. Gentili, G., Marinilli, M., Micarelli, A., Sciarrone, F.: Text categorization in an intelligent agent for filtering information on the web. IJPRAI **15**(3) (2001) 527–549