Discovering Prerequisite Relationships among Learning Objects: a Coursera-driven Approach

Carlo De Medio[1], Fabio Gasparetti[1], Carla Limongelli[1], Matteo Lombardi[3], Alessandro Marani[3], Filippo Sciarrone[1], and Marco Temperini[2]

[1] Engineering Department, Roma Tre University
Via della Vasca Navale, 79 - 00146 Roma, Italy
{limongel,gaspare,sciarro }@ing.uniroma3.it , carlo.demedio@uniroma3.it
[2] Dept. of Computer, Control and Management Engineering, Sapienza University
Via Ariosto, 25 - 00184 Roma, Italy
marte@dis.uniroma1.it
[3] School of Information and Communication Technology, Griffith University,
170 Kessels Road, Nathan, QLD, 4111 Australia
{matteo.lombardi,alessandro.marani }@griffithuni.edu.au

Abstract. In this work we address the problem of automatically finding prerequisite relations among learning materials in order to help instructional designers to speed up the course building process. Ours is a data-driven approach, where a (machine) learner is trained to classify predecessor/successor relationships, given two didactic materials in a textual form. As the training set we use the learning materials extracted from Coursera. A first evaluation shows promising results.

Keywords: Wikipedia, Learning Object, Curriculum Sequencing, Data Mining

1 Motivations, Goals and Related Work

Nowadays, Instructional Designers (IDs) can benefit of a huge source of learning materials from the Internet for the construction of Learning Units (LUs). Several Instructional Systems Design (ISD) models such as [2], [4], or [1] have been proposed to speed up and manage the process of arranging courses, but all these models require the ID to accomplish two main heavy steps: LUs building and LUs sequencing. The high availability of freely reusable LUs, however, allows the IDs to lighten the LU building task, so to focus on the sequencing problem. In such a context, uncovering educational relationships between two given LUs is a task of growing significance, allowing for a correct sequencing of the LUs for a new course. Our work addresses just this problem: given two LUs, reduced to textual form, to check whether a relationship of pre-requisite can exist between them. To accomplish this task, we followed a classic Machine Learning approach running on the DAJEE dataset [5], a dataset composed by the video transcripts of *Coursera* on-line courses. First, after having stemmed the transcripts and aggregated them by concept, we extracted some relevant features. Secondly, we annotated each set of transcripts, pertaining the same concept, by means of the *Wikipedia Miner Toolkit*. As a result we obtained, for each set of transcripts, a set of Wikipedia web pages, pertaining the same concept [3, 9, 11, 8, 10]. Then, we trained three binary learners: a decision tree, a naive-bayes and a multi-layer perceptron to inference whether a didactic relationship between the two LUs, both given in input, does exist or not. The problem of LUs sequencing has been widely addressed in literature. In [12] the sequencing engine is based on learner's current knowledge state and learning styles. Wikipedia offers a huge amount of open learning contents. Links, categories and information in templates provide structured content that can be retrieved from raw XML dumps. This makes it attractive for various research activities, such as natural language analysis, processing and translation, and it is a source of inspiration for educational activities (e.g.: [6]). Coursera is one of the largest platforms which hosts MOOCs, and DAJEE [5] is a MySql DataBase[4] built from the crawling of MOOCs hosted on Coursera. The dataset stores the URLs of the resources in Coursera, with information about i) the resources, ii) the courses where they have been delivered, and iii) the instructors who delivered them on Coursera. The resources delivered on Coursera are mostly videos. Regarding the identification of pre-requisite relationships between LUs, we found some correspondence in [13] and [14]. Recently, in [7], an early attempt to exploit Wikipedia as a source of learning materials has been proposed.

2 The Relationship Uncovering Process

In very short terms we try and associate each Coursera Education Resource (CER), and the related Concept, to a WikiPedia Topic (a web page), and then map back the conceptual pre-requisite relationships holding between two topics, onto the associated CERs and concepts. The *Relationship Uncovering Process* goes as follows.
1. The CERs are grouped conceptwise.
2. On each group a procedure of text content extraction is performed. In this way each concept C is associated to the (overall) transcript of the associated CER(s): $txt(C)$.
3. A process of annotation is performed on each $txt(C)$. Here when a *binding* between a concept C (represented by $txt(C)$) and a Wikipedia topic T_C is verified through Wikipedia Miner, we say that $Binding(C, T_C) = TRUE$.
4. Then the process of features extraction takes place on the *Binding* database: for each couple $< C, C' >$ of concepts the *actual* feature values are computed and stored. Here also the *Expected* feature values are stored for each couple. So the *Instance* database is the coupling point between the Relationships Uncovering Process, described here, and the analysis/evaluation stage, described in the next section. Moreover, we observe that:
1. Given two topics, T_C and $T_{C'}$, when the former is more general (less specific) than the latter, it is also more likely to contain a longer textual description than the latter.
2. When a topic T_C makes reference to other topics $\{T_{C'}, T_{C'\prime}, \ldots\}$ at the same time, we may well hypothesize that C is more general a concept than the $\{C', C'\prime, \ldots\}$.
3. The occurrence of concepts can be determined by the nouns occurring in the topic extracted by a Part-of-speech tagger.
4. Considering the number of words in the first sections (description) of T_C and $T_{C'}$, if the former is much greater than the latter, and there are intersections (on nouns and

[4] DAJEE can be accessed publicly for research purposes only, following the authors' approval. Apply for it by filling in the form at http://144.6.235.142/dajee

links) then it can be inferred that the CER(s) associated to T_C is a pre-requisite of the CER(s) associated to $T_{C'}$.

According to the previous observations, given a concept c and a set of related topics T_c, we define the following features: average length of the topics, number of links in the first section, average number of links in T_c, number of distinct nouns in T_c, cardinality of the intersection between the sets of nouns appearing in the topics of T_C and $T_{C'}$, average word_length of the first sections of the topics in T_c, measure of how the words used in the links from T_C' are corresponding to the nouns in T_C.

3 Evaluation

Several classifiers were trained to select the best one, i.e., the one showing the highest performance values in classifying the didactic relationships. For our experiments, we used the following classifiers: Decision trees, Multilayer Perceptron and Naive Bayes. In particular, we run two different supervised experiments, in order to verify the concept domain independence of the trained machines. The training set was taken by the domain of *Philosophy* and tested on the *Machine Learning* domain. The training set was formed by a set of couples of LUs together with their related binary outputs, as usually done in a supervised experiment. The binary outputs were set to YES or NO, standing YES for the existence of prerequisite relationship between the two LUs and NO for its absence. The results are shown in Tab. 1 for the prerequisite relationship presence and in Tab. 2 for the relationship absence, using the classic classification parameters: Recall, Precision, F_1 measure and the K statistics. The results show the multi-layer perceptron as the most promising (machine) learner. It was composed by 15 input neurons, one for each feature explained in Section 2, 2 binary output neurons and 8 hidden neurons, with one hidden layer. All these results strengthen our expectation, i.e., it will be possible to obtain a general machine learner, able to generalize in different knowledge domains to help a teacher irrespectively of the course domain as well.

Table 1. The results of the Test. Training and testing for discovering prerequisite relationship between two LUs.

Classifier	**Precision**	**Recall**	F_1	K
Naive Bayes	0.6	0.977	0.743	0.6
Decision Tree	0.335	0.974	0.498	0.004
Multi-Layer Perceptron	0.792	0.977	0.875	0.81

In this paper we proposed a novel approach to the discovery of didactic relationships among learning materials, expressed in textual form. We presented a data-driven Machine Learning approach, where, given two LUs in textual format, a binary classification (YES/NO) is produced, stating if a relationship of pre-requisite could exist. As the training set we used a subset of the courses in the Coursera repository and finally, different learners were trained and tested on different test sets with promising results.As of future work, we plan to test our approach on a broader test set and on different learning domains.

Table 2. The results of the test. Training and testing for discovering the relationship absence between two LUs.

Classifier	**Precision**	**Recall**	F_1	K
Naive Bayes	0.983	0.67	0.797	0.6
Decision Tree	0.708	0.032	0.061	0.004
Multi-Layer Perceptron	0.987	0.872	0.926	0.81

References

1. Allen, M., Sites, R.: Leaving ADDIE for SAM: An agile model for developing the best learning experiences. American Society for Training and Development (2012)
2. Branch, R.M.: Instructional Design: The ADDIE Approach. Springer Publishing Company, Incorporated, 1st edn. (2009)
3. De Medio, C., Gasparetti, F., Limongelli, C., Sciarrone, F., , Temperini, M.: Automatic extraction of prerequisites among learning objects using wikipedia-based content analysis. In: Proc. of the 13th Int. Conf. on Intelligent Tutoring Systems. ITS '16, Springer (2016)
4. Dick, W., Carey, L.and Carey, J.O.: The systematic design of instruction. Upper Saddle River, N.J: Merrill/Pearson (2009)
5. Estivill-Castro, V., Limongelli, C., Lombardi, M., Marani, A.: Dajee: A dataset of joint educational entities for information retrieval in technology enhanced learning. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 681–684. ACM (2016)
6. Forte, A., Bruckman, A.: From wikipedia to the classroom: Exploring online publication and learning. In: Proceedings of the 7th International Conference on Learning Sciences. pp. 182–188. ICLS '06, International Society of the Learning Sciences (2006)
7. Gasparetti, F., Limongelli, C., Sciarrone, F.: Wiki course builder: A system for retrieving and sequencing didactic materials from wikipedia. In: Information Technology Based Higher Education and Training (ITHET), 2015 International Conference on. pp. 1–6 (June 2015)
8. Gentili, G., Marinilli, M., Micarelli, A., Sciarrone, F.: Text categorization in an intelligent agent for filtering information on the web. IJPRAI 15(3), 527–549 (2001)
9. Limongelli, C., Lombardi, M., Marani, A., Sciarrone, F.: A teaching-style based social network for didactic building and sharing. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7926 LNAI, 774–777 (2013)
10. 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 2(1-2), 58–82 (2015)
11. Limongelli, C., Mosiello, G., Panzieri, S., Sciarrone, F.: Virtual industrial training: Joining innovative interfaces with plant modeling. In: Proceedings of the International Conference on Information Technology Based Higher Education and Training, ITHET 2012 (2012)
12. Limongelli, C., Sciarrone, F., Temperini, M., Vaste, G.: Adaptive learning with the ls-plan system: A field evaluation. IEEE Trans. on Learning Technologies 2(3), 203–215 (2009)
13. Scheines, R., Silver, E., Goldin, I.: Discovering prerequisite relationships among knowledge components. In: Stamper, J., Pardos, Z., Mavrikis, M., McLaren, B. (eds.) Proceedings of the 7th International Conference on Educational Data Mining. pp. 355–356. ELRA (May 2014)
14. Vuong, A., Nixon, T., Towle, B.: A method for finding prerequisites within a curriculum. In: Pechenizkiy, M., Calders, T., Conati, C., Ventura, S., Romero, C., J. Stamper, J. (eds.) The 4th International Conference on Educational Data Mining (EDM 2011). pp. 211–216 (2011)