An Intelligent Agent with Ontological Knowledge: Classification of Educational Materials to Support the Creation of Online Courses

Carlo De Medio

Engineering Department, Roma Tre University,
Via della Vasca Navale, 79 - 00146 Roma, Italy
carlo.demedio@uniroma3.it

Abstract. The composition of a course for e -learning platforms, through the selection and the sequencing of teaching materials (Learning Object) is a complicated process and is made manually by teachers. Target of this research is, through the study of the dependencies between topics related to the LO in Wikipedia, understanding the prerequisite and successor dependencies between materials in order to create an intelligent tutor that can assist teachers in the creation of a course, including the process of searching the materials, to their automatic or semi-automatic sequencing in order to optimize the learning process by students.

Keywords: Sequencing, E-Learning, Wikipedia

1 Motivation of the research

In the age of e-learning, many instructors are faced the hard task of building a web-based course; a lot of different platforms are growing like, public and private university and corporation platforms. The majority of these platforms focus their aim on helping both the teachers to recover materials for the creation of courses from the web and the students to benefit from this. The primary target of many existing systems is to share the knowledge, through the repository of Learning Objects on the web (Ariadne, Merlot, CNX, Wisk..). In those repository there isn't correlation between materials, and the reuse of those knowledge, based on a meta-classifications tagged manually by creator in different formats, is almost impossible. This information are also subjective and hard to interpretate by other teachers because the non existence of a standar de facto. Furthermore the sequencing process of those LO is strongly influenced by the teacher pref- erences. Specifically, the current systems are unable to overcome the manual classifcation limit; its not possible, starting from a concept, to generate automatically a chain of relations between concepts, and from those chain to obtain useful hints for the completion of the course. Given two Learning Objects LO1 and LO2 the dependency relation of prerequisite and successor is defined as:

1. prerequisite: $LO_1 > LO_2$
2. successor: $LO_1 < LO_2$

This research seeks to overcome these deficiencies of current systems, dealing with two key points:

- The study through shared knowledge bases (eg. Wikipedia) of the prerequisite and successor dependency relation between two generic LOs (e.g. html page, simple text ..), in order to obtain an intelligent agent able to generate chains of linked concepts and create a navigable graph of knowledge.
- The implementation of a virtual tutor that using the agent, is able to assist the teacher in the complete process of creating a course in two ways:
  1. Assisted: the teacher inserts an initial concept and recive back a list of possible next LOs; a selection in this list corresponds to the generation of a new list of concepts related to the one selected until the completion of the course.
  2. Automatic: the teacher inserts an initial concept and the tutor automatically creates a course according to teacher styles. He can accept it all or change any part of the course, until the process is complete; all the interactions are traced by the teacher model.

The goal of the final system is the creation of what is described above on a platform based on user models (student and teacher) in order to keep track of past decisions taken by individual users and weigh that decision with the tutor decisions in order to optimize the creation and the fruition of the courses.

2 Related work

The association of teaching materials from different sources within a course is a hard task and can not be treated as a simple additive process; at each iteration all the historic matured must be re-evaluate in order to optimize the process. A popular method is the integration of user models based on Teaching Styles; as is shown in [1] one of the most important task is to undestrand how the interaction with the system has to modify the user models. The research in the field of sequencing applied to e-learning aims to automate this process; creating a custom sequence of activities on demand based on user models and optimize experience for each user [2]. An example of a platform based on user models and learning style is $MoodleLS$ reported in [5][6] where the main theme is to get a dynamic sequencing of the course for each student and the research through an initial test of knowledge of the student before the course. According to various interactions with the system and the evolution of the model the sequencing may change and/or add concepts necessary to the completion of the course. To evaluate the content of the materials were used solutions through the Web, the most widely used are Wikipedia (exemple of approch [3]) in which a large part of the materials are inserted by teachers, as reported in the paper [4]. However, Wikipedia is a very wide base of knowledge and is not enough meta-dated, the interpretation

is subjective by the community. The advantage of Wikipedia is the materials homogeneity that easily allows to automate the process of knowledge extraction [13]. The most common techniques adopted for content indexing are Dublin Core [10], IMS Metadata (defined by IMS Global Learning Consortium [10]) and IEEE LOM (Learning Object Metadata defined by Learning Technology Standardization Committee, LTSC dell IEEE, [11]); before this techniques is fundamental to pre-process the materials to obtain part- of-speech tags and low-level syntactic features [12]. Some systems have tried to improve the search of LOs with a concept based ontology, beacause an ontology facilitates the sharing and reuse of knowledge [7]. In [14], is presented an early attempt to exploit Wikipedia as a source of learning materials based on the TFxIDF text distance. Another great source of knowledge is Coursera, the biggest platform that host MOOC; an interesting project is the DAJEE database [15] that contains all Courseras structure obtained by crowling and easily accessible.

3 Main research questions and description of the undertaken research

The first question is how to recognize the prerequisite relationship between two generic LOs, unstructured and coming from different sources; the research for this problem looks through an annotation service in Wikipedia (and also Tagme) of the topics associated with the two LOs. Then, through the study of the relations between those topics the goal of the research is to assert the existence of the prerequisite/successor relationship. Starting from the work of [8], the research aims to extend the sets of features associated to the two pool of topics to be used in a machine learning algorithm (as training) and generates an intelligent agent able to solve the proposed problem. The prosses should:

- get topics from LOs through Wikipedia Miner [19] and Tagme [18],
- extract the features from this pool of topics associated to the LO,
- train and test the machine learning algorithm with those data,
- evaluate the classification process by the main measures in the literature.

In picture [1] is reported the process schema.

3.1 Tagme and Wikipedia Miner services

The services offered by Wikipedia miner and Tagme are similar. The services take as input a simple text; after a stemming process the text is associated to a set of "Tags" that relate to wikipedia relevant pages. The output is a json file containing the information of the identified Tag (e.g. id wikipedia page, title, categories associated ....) and a probabilistic measure of how much the associated article is relevant in the given context.

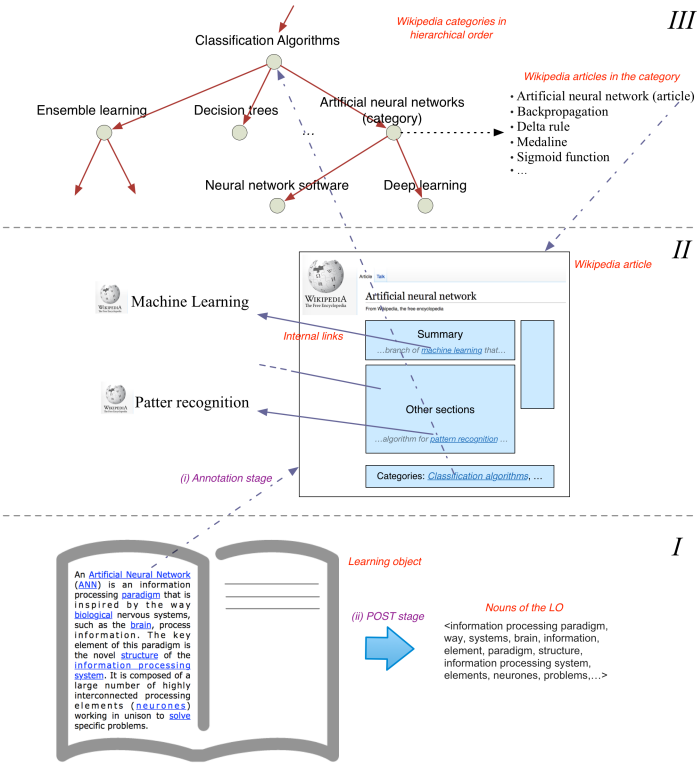

Fig. 1. Functional Architecture of the Relationships Uncovering Process

3.2 system validation

Furthermore we must verify if the intelligent agent will succeed, through the training on various arguments to be domain independent and therefore useful in any field, even where it still has not been trained. For this purpose, the dataset will be split into macro categories and will be used in various tests using a one domain out evaluation; the goal will be to maximize the results of precision, recall, f-masure, roc-area and accuracy. The research will seek to assert if the decisions taken by the teachers are able to improve the ontological decision-making process answering the question: if teachers with similar profiles prefer similar materials [1]. To this purpose, once the system will be implemented, it will require a uniform testing process. We will be considered two groups of user:

1. a group that will use the system to generate courses,
2. a group who will use the system with the user models.

Finally, the courses will be followed by a pool of students and the results collected for evaluation.

4 Relation Recognizer

The Recognizer is composed of two binary classifiers able to recognize relationships $LO1 > LO2$ and $LO2 < LO2$; is based on a set of 19 features for each pair of Learning Object in input and trained through a Machine Learning algorithm.

4.1 Classifer Features

Features associated to the *LearingObjects*

- set of nouns in $LO's$,
- set of Wikipedia articles annotated to $LO's$,
- length in terms of number of words,

Features associated with a Wikipedia article c or category k related to the $LO's$

- length in terms of number of words.
- length in terms of number of words of the summary section.
- average length of the articles associated to the $LO's$
- average length of the summary section of the articles associated to the $LO's$
- set of links in a LO to other articles.
- title of the articles associated to a LO.
- Wikipedia categories assigned to a LO.
- set of Wikipedia articles in the Wikipedia category assigned to a LO.

Features associated with a pair of $LO's$

- set of link in LO_1 that contains in the referenced text a LO_2 nouns and viceversa
- set of nouns that belong to each $LO's$
- the union of the set of nouns of LO_1 and LO_2
- average length of the summary section of the articles associated to the $LO's$
- set of links in a LO to other articles.
- set of link of LO_1 that points to LO_2 and viceversa.
- Wikipedia categories assigned to a LO.
- counts the number of super-categories or sub-categories that a LO has in common with the other one at distance d.

4.2 Classifier Evaluetion

The actual evaluation is made on CrowdComp MTurkData dataset, already treated in [9], and a set of couses from Edx and Udacity; For those courses taken from on-line repository a pool of domain expert (e.g. Teachers from Roma TRE, Teachers from Sapienza, ...) created the dependency list for the evaluetion process. The CrowdComp MTurkData dataset consists of five domains, with a total amount of 206 prerequisites and 1600 LOs. To our knowledge, it is the only public dataset that provides enough depth for including different topics

and a sufficient amount of prerequisites. It provides the text conent of each LO, which have been collected from a real-world collection of learning material. The Amazon Mechanical Turk [17] crowdsourcing platform has been exploited for recruiting participants that manually defined the prerequisites relationships. The Courses statistic are reported in Table [1]

Table 1. Statistics about the datasets.

	ID	*Domain*	*LOs and courses*	*Prerequisites*
CrowdComp	1.	Meiosis	400	67
	2.	Public-key Cryp.	200	27
	3.	Parallel Postulate	200	25
	4.	Newton's Laws	400	44
	5.	Global Warming	400	43
Udacity	6.	Biology	206 (1)	16
	7.	Computer Science	2,396 (4)	68
	8.	Math, Statistics & Data Analysis	1,759 (3)	12
	9.	Physics	546 (1)	10
	10.	Psychology	690 (1)	26
edX	11.	Design	66 (2)	10
	12.	Economy and Finance	91 (2)	12
	13.	Engineering & Project Management	582 (15)	64
	14.	Politics	62 (2)	8

All this data were evalueted using a Naive Bayes classifier, a Tree based classifier and a Neural Network; the results using the technique of one domain out using the tool Weka [16] are shown in the Table [2].

Table 2. Performance outcomes. Standard deviation σ over the courses inside the parentheses.

	Pr	Re	F1	A	AUC
0-RL	0.34 (0.01)	0.58 (0.01)	0.42 (0.01)	0.68 (0.09)	0.50 (0.00)
C4.5	0.78 (0.01)	0.74 (0.02)	0.74 (0.02)	0.74 (0.02)	0.74 (0.02)
MLP	**0.81** (0.02)	**0.78** (0.03)	**0.78** (0.03)	**0.78** (0.03)	**0.87** (0.01)
NB	0.71 (0.04)	0.70 (0.03)	0.69 (0.03)	0.70 (0.03)	0.78 (0.02)

As baseline approach, a *Zero Rule classification* (0-RL) has been considered, which relies on the frequency of targets and predicts the majority target category. the MLP proves its ability to learn complicated multidimensional mapping going beyond traditional regression and Bayesian approaches, which obtain similar performances. In particular, the Naive Bayes approach shows better results in

terms of AUC, but the C4.5 decision tree improves the precision and recall of the classification.

5 Advancements that can be derived in the field of interest,by the proposed research

As mentioned, the theme of e-learning is a consolidated reality in the world. In the era of adult learning, just-in-time learning or life-long learning the results of research are proposed to solve beyond classic sequencing problems and reuse of materials and courses; providing teachers with an assistant to facilitate the task of setting up courses and optimize their usability. It also makes possible for any person, not domain expert, through the virtual tutor create a self-build path using the knowledge of teachers who have already used the platform. Ideally, after the cold start, each student could have his own program of study configured by himself, and calibrated transparently by the virtual tutor based on its user model. Finally, the capability of the intelligent agent in recognizing through Wikipedia relations between concepts, is applicable outside the scope of the research, in particular:

- inside a semantic analysis system to calculate the nearness between concepts and their relationships in order to solve the problem of disambiguation,
- in the field of knowledge management in order to automatically classify knowledge in macro-categories,
- extending prerequisite relation from Learning Object to courses to gerenate a curriculum for the student.

References

1. A. Grasha, Teaching with style:Apratical guide to enhancing learning by understanding teaching and learning styles (1996)
2. Brusilovsky P., Vassileva J.: Course Sequencing Techniques for Large-Scale Web-based Education. Int. J. of Continuing Engineering Education and Lifelong Learning 13, 7594 (2003)
3. 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. 16 (June 2015)
4. K. Parker and J. Chao, Wiki as a Teaching Tool. Interdisciplinary Journal of E-Learning and Learning Objects. 5772 (2007)
5. C. Limongelli, M. Lombardi, A. Marani, F. Sciarrone, M. Temperini, A recommendation module to help teachers build courses through the Moodle Learning Management System, New Review of Hypermedia and Multimedia, 5882 (2015)
6. 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), 203215 (2009)
7. V. Devedzic, Education and the Semantic Web, International Journal of Artificial Intelligence in Education 14 (2004)

8. F. Gasparetti, C. Limongelli, F. Sciarrone, Exploiting wikipedia for discover- ing prerequisite relationships among learning objects, in: 2015 International Conference on Information Technology Based Higher Education and Training, ITHET 2015, Lisbon, Portugal, June 11-13, 2015, IEEE, 2015, pp. 16. doi: 10.1109/ITHET.2015.7218038. URL http://dx.doi.org/10.1109/ITHET.2015.7218038
9. C. Liang, Z. Wu, W. Huang, C. L. Giles, Measuring prerequisite relations among concepts, in: Proceedings of the 2015 Conference on Empirical Methods in Nat- ural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 16681674. URL http://aclweb.org/anthology/D15-1193
10. W. Dick, J. O. Carey, L.and Carey, The systematic design of instruction, Upper 340 Saddle River, N.J: Merrill/Pearson, 2009.
11. A. H. Brown, T. D. Green, The essentials of instructional design: Connecting 345 fundamental principles with process and practice, Routledge, 2015.
12. D. Roy, S. Sarkar, S. Ghose, Automatic extraction of pedagogic metadata from learning content, Int. J. Artif. Intell. Ed. 18 (2) (2008) 97118. URL http://dl.acm.org/citation.cfm?id=1454082.1454084
13. Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F.r., 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)
14. 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. 16 (June 2015)
15. Estivill-Castro, V., Limongelli, C., Lombardi, M., Marani, A.: Dajee: A dataset of joint educational entities for information retrieval in technology enhanced learning. In: In Proceedings of the 39th International ACM SIGIR Conference. ACM (2016)
16. University of Waikato: wikipedia miner information, https://github.com/dnmilne/wikipediaminer/wiki/About-wikipedia-miner, last accessed 2016.06.23
17. Amazon.com, Mechanical turk, last visited on 31 August 2016. URL https://www.mturk.com/mturk/
18. P. Ferragina, U. Scaiella, Fast and accurate annotation of short texts with wikipedia pages, IEEE Softw. 29 (1) (2012) 7075. doi:10.1109/MS.2011. 122. URL http://dx.doi.org/10.1109/MS.2011.122
19. W. Project, Mediawiki, last visited on 31 August 2016. URL https://$www.mediawiki.org/wiki/API : Main_page$