A Content-Based Approach for Supporting Teachers in Discovering Dependency Relationships Between Instructional Units in Distance Learning Environments

Fabio Gasparetti^(✉), Carla Limongelli, and Filippo Sciarrone

Roma Tre University, Via Della Vasca Navale 79, 00146 Rome, Italy
{`gaspare, limongel, sciarro`}@dia.uniroma3.it

Abstract. eLearning courses are usually built in such a way that training resources are organized for making the learning process effective. In most cases, the top-level learning objective will have prerequisites which must be satisfied. Those prerequisites should be formally identified by a hierarchy of dependencies built accordingly. We evaluate a series of hypotheses for understanding the feasibility of automating this task by means of a general-purpose content-based approach that exploits semantic analysis techniques.

Keywords: Sequencing · Distance learning · Semantic analysis

1 Introduction

Research on Course sequencing aims to automatically produce a personalized sequence of didactic material and activities on the basis of each student's needs and preferences [3].

Still, with advances in science and technology a wide range of technological options are available to the teachers, and distance education offers us big opportunities in comparison with traditional classroom instruction, reaching a wider student audience, meeting the needs of students who are unable to attend on-campus classes and linking students from different social, cultural, economic, and experiential backgrounds. Nevertheless, teaching and learning at a distance is as effective as traditional instruction when the employed learning material, method and technologies are appropriate to the instructional tasks.

Standalone instructional units (or *learning objects*), where each unit covers a single learning point in its entirety, and large online repositories (e.g., Connexion¹ and Ariadne²) simplify the educators to share, manage and use educational resources. Additional resources can also be retrieved by selectively crawling the web [9]. Once a subset of these objects have been selected, instructional designers

¹ <http://cnx.org>.

² <http://www.ariadne-eu.org>.

must decide the sequence in which learners will accomplish these objectives. It is not a trivial task and involves a variety of skills.

The mutual-dependency relationship between the content of the learning object and the context into which it is being placed determines whether or not the object “fit” into that context [4, 7, 8]. The correct concatenation or *sequencing* of contextualized educational resources develops the required meaningful distance learning course.

The position paper reports an original general-purpose content-based approach for assisting teachers in defining *prerequisite* relationships between text-based learning objects and, at the same time, providing a list of valid sequences of learning objects that satisfy them. Semantic analysis techniques identify relevant Wikipedia concepts mentioned in each learning object. After this annotation step, by tapping the vast semi-structured Wikipedia knowledge, the constraints between learning objects are identified. Because a large set of reusable digital resources include text and web pages, Microsoft word and pdf file formats, the proposed approach is focused on the learning objects which more easily can be assigned a sort of keyword-based representation of their content. As part of an ongoing research, we are currently collecting evaluation test results to prove the validity of the approach.

2 Prerequisite Extraction

Given a pair of learning objects, lo_i and lo_j , the relationship $lo_i \rightarrow lo_j$ specifies the *ordering constraint* under which the topic containing lo_j is sequenced first. The set of those conceptual prerequisite constraints defines how two concepts are related to each other and can be used to derive instructional design rules that must be validated against the course, e.g., “all topics covering broader concepts are sequenced before those more specialized”.

Each learning object can include a variety of different topics. For this reason, two steps are required: (1) the identification of a knowledge base that contains a taxonomy of the topics and (2) a methodology for assigning to each object a subset of those topics.

Because manual ontology engineering requires great effort, especially if the ontology must be kept up to date with future topics, we try to overcome such problems by relying on Wikipedia, a wide coverage multilingual online encyclopedia developed by a large number of users. Each published article p can belong to one or more categories $C_p = \{c_1, c_2, \dots, c_n\}$, subset of available ones on Wikipedia $C_p \subset C_w$. The C_p categories describe the domains related to the article. To organize Wikipedia for easy access to pages, contributors are given guidelines for categorizing articles and naming new categories.

Categories themselves may have subcategories, or reversely, supercategories. However, the categories form that weak conceptually-related thesaurus including a variety of semantic similarity relationships. On the other hand, prerequisites are mostly based on hierarchical relationships, such as IS-A, hypernym-hyponym, part-whole et cetera. Other taxonomies that explicitly include those kinds of

relationships (e.g., Wordnet) are not frequently updated or more likely fail to include technical or specific concepts discussed in high education courses.

Given two learning objects lo_i and lo_j , we run the Wikipedia Miner toolkit³ for detecting and disambiguating the Wikipedia topics when they are mentioned in the textual content extracted from lo_i and lo_j , obtaining two subsets of Wikipedia articles, respectively.

For studying the potential correlation between the prerequisite relationships of pairs of LOs and the two sets of Wikipedia articles describing the concepts covered by the LOs we formulate three hypotheses.

The first one is structural-based. Given two learning objects lo_i and lo_j , and the related subsets of Wikipedia articles, we can see that each article is assigned to one or more categories. We can obtain the two sets of categories associated with the two learning objects, C_{lo_i} and C_{lo_j} , respectively. Given two categories $c_i \in C_{lo_i}$ and $c_j \in C_{lo_j}$, we denote with $c_i \rightarrow_w c_j$ the shortest direct path through the Wikipedia taxonomy connecting c_i to c_j , where c_i is a more general category. The function $d : (C_w, C_w) \rightarrow \mathbb{Z}$ returns the shortest length of $c_i \rightarrow_w c_j$, ∞ if it does not exist.

The first hypothesis **H1** states that $lo_i \rightarrow lo_j$ exists if:

$$\exists c [d(c, c_{lo_i}) < d(c, c_{lo_j}) \wedge d(c, c_{lo_j}) < \infty \wedge \frac{\max(d(c, c_{lo_i}), d(c, c_{lo_j}))}{d(c, c_{lo_i}) + d(c, c_{lo_j})} > \alpha] \quad (1)$$

where $\alpha \in (0, 1]$ is a constant and $c_{lo_i} \in C_{lo_i} - (C_{lo_i} \cap C_{lo_j})$, $c_{lo_j} \in C_{lo_j} - (C_{lo_i} \cap C_{lo_j})$. In summary, given two distinct categories, each from one of two learning objects, if the distance that connects them, even through a common ancestor category, is less than a given threshold, the prerequisite relation exists. The order of the two learning objects is determined by analyzing how deep the c_j category is placed in the taxonomy in comparison with the category c_i .

The other two hypotheses are content-based. The hypothesis **H2** states that if lo_j has been associated with an article that contains a link to a page associated with lo_i at the beginning of the description, $lo_i \rightarrow lo_j$ exists. If the authors have found the necessity to define a concept by means of other concepts, probably the former one is more specific than the latter. The last hypothesis **H3** states that if the Wikipedia associated with lo_i is substantially longer than the one of lo_j , $lo_i \rightarrow lo_j$ exists. The rationale is that more general topics need longer discussions to be described than a very specific one.

3 Early Evaluation

We briefly give account to the criteria we are currently following to assess the appropriateness of the proposed approach. The research question of this study is to observe whether there is statistical relationship between the criteria for each hypothesis we introduced and the presence of a prerequisite relationship.

³ <http://wikipedia-miner.cms.waikato.ac.nz>.

Table 1. Average precision, recall and F1 measures for the evaluated hypotheses.

	Precision	Recall	F1
H0	0.394	0.231	0.291
H1	0.501	0.363	0.421
H2	0.493	0.041	0.075
H3	0.540	0.300	0.386

The analyzed data is obtained from a large set real online courses that cover various different topics. The prerequisite relationships are identified by the help of a group of teachers. We analyzed four computer science courses.

For estimating the effect of each criterion, we are currently making use of traditional Information Retrieval measures, namely, Precision, Recall and F1-measure. Table 1 reports the average values for every hypothesis over the four courses. . H0 hypothesis represents the baseline that, for each pair $\langle lo_i, lo_j \rangle$ randomly assigns the $lo_i \rightarrow lo_j$ relationship. In spite of the few input courses considered, the outcomes shows significantly deviations from a random approach that does not consider the content of the LOs.

The low recall of H2 hypothesis can be partially justified by the wrong annotations provided by the Wikipedia miner tool. By manual editing the annotations with the correct Wikipedia articles associated with the learning objects, the hypothesis reached similar recall values in comparison with the other ones.

4 Related Work

To the best of our knowledge, our work makes the first attempt to exploit weak-semantic taxonomies such as Wikipedia for the sequencing of learning objects.

Scheines *et al.* [10] use casual models search to identify prerequisites among knowledge components represented as latent variables. The approach does not analyze the content of learning objects but exploits test data collected measuring the student skills after attending the courses. Similarly, Voung *et al.* [11] propose to analyze large-scale assessment to determine the dependency relationships between knowledge units.

5 Conclusion and Future Work

Preliminary attempts to identify prerequisite relationships between learning objects have been done through the discussion of three hypotheses. We are currently concluding the experimental evaluation of the proposed approach. Early preliminary results on four courses are promising.

As for the future work, learning objects might cover particular skills that do not have compulsorily to be labeled to Wikipedia articles. For example, both “recognize task bars” and “dependency markers” depend on the “uses of Gantt

charts”. While all those learning objects will likely cover similar topics, prerequisites exists though. Content-based approaches might fail to find distinctive elements to find those relationships. One possible solution is the combination of multiple taxonomies and web corpora that can be analyzed in order to find valuable relationships, which can then be included in the prerequisite identification [2]. If two learning objects reference one another, a single representation of the covered topics can be defined by analyzing the learning path and the underneath motivations that the author of the didactic material has considered through those citations [5].

As for the Wikipedia taxonomy, different types of Wikipedia categories exists, each one based on the type of information they encode. Assuming that the IS-A and similar relationships are the only ones in Wikipedia is inaccurate and can negatively affect the precision of the inference. Further examinations are required.

Once the prerequisite identification has been correctly validated, the recommendation of learning objects can be extended by querying the learning repositories in search of new didactic material that has not been previously included by the instructor, e.g., learning objects covering similar topics. By profiling current users interests and their reactions to the retrieved objects (see for example recent advances on information sharing in social environments introduced in [1,6]), clusters of similar users can be identified in order to share among them objects that results in the maximization of the preference satisfaction and performance assessments.

References

1. Arru, G., Feltoni Gurini, D., Gasparetti, F., Micarelli, A., Sansonetti, G.: Signal-based user recommendation on twitter. In: Proceedings of the 22nd International Conference on World Wide Web Companion, pp. 941–944. WWW 2013 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2013). <http://dl.acm.org/citation.cfm?id=2487788.2488088>
2. Biancalana, C., Gasparetti, F., Micarelli, A., Sansonetti, G.: Social semantic query expansion. *ACM Trans. Intell. Syst. Technol.* **4**(4), 60:1–60:43 (2013). <http://doi.acm.org/10.1145/2508037.2508041>
3. Brusilovsky, P., Vassileva, J.: Course sequencing techniques for large-scale web-based education. *Int. J. Contin. Eng. Edu. Lifelong Learn.* **13**, 75–94 (2003)
4. De Marsico, M., Sterbini, A., Temperini, M.: A framework to support social-collaborative personalized e-learning. In: Kurosu, M. (ed.) *HCI/HCI 2013, Part II*. LNCS, vol. 8005, pp. 351–360. Springer, Heidelberg (2013)
5. Gasparetti, F., Micarelli, A., Sansonetti, G.: Mining navigation histories for user need recognition. In: Stephanidis, C. (ed.) *HCI 2014, Part I*. CCIS, vol. 434, pp. 169–173. Springer, Heidelberg (2014)
6. Gurini, D.F., Gasparetti, F., Micarelli, A., Sansonetti, G.: iSCUR: interest and sentiment-based community detection for user recommendation on Twitter. In: Dimitrova, V., Kuflik, T., Chin, D., Ricci, F., Dolog, P., Houben, G.-J. (eds.) *UMAP 2014*. LNCS, vol. 8538, pp. 314–319. Springer, Heidelberg (2014)

7. Limongelli, C., Lombardi, M., Marani, A., Sciarrone, F.: A teacher model to speed up the process of building courses. In: Kurosu, M. (ed.) *HCI/HCI 2013, Part II*. LNCS, vol. 8005, pp. 434–443. Springer, Heidelberg (2013)
8. Limongelli, C., Sciarrone, F., Vaste, G.: LS-PLAN: An effective combination of dynamic courseware generation and learning styles in web-based education. In: Nejdl, W., Kay, J., Pu, P., Herder, E. (eds.) *AH 2008*. LNCS, vol. 5149, pp. 133–142. Springer, Heidelberg (2008)
9. Micarelli, A., Gasparetti, F.: Adaptive focused crawling. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) *Adaptive Web 2007*. LNCS, vol. 4321, pp. 231–262. Springer, Heidelberg (2007)
10. 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. European Language Resources Association (ELRA), May 2014
11. 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)