Automatic Learning Content Sequence via Linked Open Data

Ruben Manrique

Systems and Computing Engineering Department, School of Engineering,
Universidad de los Andes, Bogotá, Colombia
rf.manrique@uniandes.edu.co

1 Problem Statement

The paradigm of lifelong learning supported by technology is redefining the classic approach making closed models in which objectives, content and sequence are predetermined more open and self-directed. In the new era of learning, people take on learning processes that are not necessarily carried out under the guidance of a tutor. In this format, learners are in charge of searching for and selecting the most appropriate resources to accomplish their goals. Learning content is any type of resource relevant to a learning goal, and the main source for these is the Web. However, even though most documents are not annotated as educational on the Web, they are still being used as learning resources. Given these characteristics, self-directed learners face different challenges in selecting and organizing adequate learning resources. We identify three main challenges: effective search and selection of educational content, sound organization of this content, and personalization of these two previous processes. Although personalized learning is recognized as a key concept in future learning support systems, our proposal is centered on the other two challenges, mainly educational content search, selection and organization.

The first of these challenges is that learners do not always have enough literacy skills to do an effective and selective search [8]. There is a great volume of resources available on the Web that can overwhelm learners. Secondly, unlike traditional approaches, learners do not have the assistance of an expert to structure and organize the resources so that the learning goal is achieved in a logical manner. The sequence with which the knowledge is acquired is important because complex concepts require the understanding of more basic ones. Despite that considerable attention has been paid to content organization in e-learning courses, most systems are still under development [3, 17] or are implemented in closed scenarios, or in courses where objectives, learning materials and sequence are not driven by the learner himself [18, 7, 1]. This proposal addresses this problem and explores strategies to organize and select resources by automatically enriching the resources with metadata about the concepts involved and an ordering strategy. Concepts found in Knowledge Bases (KB) belonging to the Linked Open Data (LOD) initiative are used concept space. The ordering strategy can be defined explicitly via a specific property modeled in the KB ontology or can be generated by prerequisite relationships discovered automatically through the background knowledge found in such bases.

2 Relevancy

On a more practical level, we want to support lifelong learners. As a result of the use of our system, learners will spend less time searching for and selecting the learning resources that best suits their learning goals. Also, showing a plan for using materials gives learners a general overview of a topic allowing them a more comprehensive understanding and a logical order to follow.

This dissertation will also contribute to current research in Technology Enhanced Learning and in Semantic Web. Regarding Technology Enhanced Learning, our work will contribute to advance current research on connecting Linked Data with open learning resources, in particular for autonomous lifelong learning. The Semantic Web research community should benefit from our products and results as all our algorithmic proposals take advantage of background knowledge found in KBs belonging to the LOD initiative to infer knowledge about content documents and relationships between them, which can be used for purposes other than educational ones.

3 Related Work

Learning Content Sequence: There have been several works on learning content sequence in closed learning scenarios such as e-learning courses. In such scenarios, the problem is known as "pedagogical sequencing" or "curriculum sequencing" and is defined as the problem of how the learning material is arranged and delivered to a particular learner according to his/her preferences, level of knowledge and/or the pedagogical conventions [2, 18, 7, 1]. While some approaches solve the problem sequentially by defining events and rules that guide the construction of the sequence [18], others have modeled it as a constraint satisfaction problem in which diverse computational techniques can be employed [2, 23]. These approaches have proven useful in the context of well defined, closed domain ontologies and of learning objects (LO) annotated with standards like LOM, SCORM or LRMI. In these cases, the sequencing of Learning Content can be achieved by following subsumption relations in the domain ontology and by aligning ontology concepts and LO annotation terms. The main limitations of these approaches are availability, scalability and maintenance. To start, not all domains have well established and agreed on ontologies and even in the case of domains having these ontologies both tasks: LO annotation and ontology construction and maintenance require intensive expert human intervention, as they are not completely automatic tasks. A second problem refers to scope. In the context of life-long self-directed learning, the learner might be interested in cross-domain problems and pertinent content might include learning resources not originally labeled as LO. In order to address the above issues, research has been oriented to the automatic extraction of concepts and prerequisite relationships to guide the selection and organization of learning resources.

Automatic Extraction of Concept-Based Metadata: [22] developed an automatic annotation tool for annotating documents with metadata such as concepts, type of concepts, topic and the learning resource type. They justified the importance of extending IEEE LOM metadata with concept based attributes

since it allows a better retrieval process. Ontology domains are used to map lexical terms to concepts. To evaluate the importance of a concept, they analyze the frequency of its related concepts. The type of concept (i.e. outcome, prerequisite, defined and used) is identified by analyzing the documents with a shallow parsing approach and by using some inference rules. In a more recent but similar works, [9, 6] enrich learning resources metadata with semantic concepts from a domain ontology. Using an ontology extraction algorithm named Hieron, they extract the concepts covered by the content resource with an associated weight that represents the degree of pertinence. [4] addresses the problem of determining an effective learning path from a corpus of Web documents via the annotation of outcome and prerequisite concepts. They employ a machine-learning techniques to predict the class of concept on the basis of contextual and local text features. Regarding information extraction techniques that take advantage of LOD-KB background knowledge, we only found the work done by [10]. In this work, authors propose an automatic semantic description of Web Documents based on LOD concepts. The semantic description is called "fingerprint" and allows to judge if a Web resource is relevant with respect to a learning context or not. We do not find evidence of its used to select or organize resources.

Prerequisite relationships identification: The problem of identifying prerequisite relationships in open setting has been addressed by several researchers. We focus here on the state of the art proposals [24, 25, 11]. [24] proposed a supervised method for predicting prerequisite dependencies between Wikipedia pages. They used a set of features such as random walk with restart (RWR) and Pagerank scores in combination with a Maximum Entropy (MaxEnt) classifier. Later [25] formulates an optimization problem to obtain concepts maps linked with prerequisite relationships extracted from textbooks. A set of objective functions is proposed taking advantage of the order and frequency in which concepts appear in the textbook structure. Finally, a reference distance RefD is proposed in [11] to measure a prerequisite relation among concepts. Interesting this measure could be generalizable to different concept spaces and weighting functions. Our approach is different from the preceding ones for we rely on LOD-KBs (e.g. DBpedia, Yago) as Universal concept space. We believe that employing similarity measures for LOD concepts and the background knowledge present in the KB, it is possible to infer candidate prerequisite relationships.

4 Hypotheses and Research Questions

In order to address the main problem we propose the following hypotheses and research questions:

(H1) It is possible to generate a learning concept graph (LCG) that expresses prerequisite relationships from the well-structured information in LOD KB. **(Q1)** How can LOD KB background information be exploited to infer candidate prerequisite relationships? **(Q2)** How accurate are these inferred prerequisite relationships?

(H2) It is possible to augment learning resources with metadata associated with the main concepts that it addressed. **(Q3)** How can a concept based descrip-

tion for learning resources be extracted (automatically) based on its content? **(Q4)** How well does this concept-based description represent the resource?

5 Approach

This dissertation is aimed at selecting learning resources to satisfy a learning need and at automatically inferring prerequisite relationships between concepts related to this learning need that allow a logical ordering of the found resources. The resources, therefore, must follow a representation that allows to discern the main concepts that it approaches. Knowing how the concepts are related and the concepts that each resource addresses is possible to select and order the learning resources for a given learning objective.

In general, our methodology to address the above issues relies on (i) the automatic discovery of prerequisite relationships between concepts and (ii) the automatic construction of a concept-based representation for learning resources. Both processes are guided by the knowledge found in LOD-KBs and allow to enrich a learning resource with metadata about the main concepts it addresses and their corresponding prerequisite concepts.

There are two main components in our proposal: the Learning Concept Graph (LCG) and the Automatic Learning Content Sequence (ALCS) generator service. LCG infers automatically prerequisite relationships among concepts by analyzing their semantic relationships and linkages. This information is exploited by ALCS to extract the concepts and the order in which they have to be cover to achieve an input-learning goal. Finally, by extracting information about the main concepts involved in a learning resource, the ordered sequence of concepts is transformed into an ordered sequence of resources. The paragraphs below explain both components.

Learning Concept Graph (LCG): The main objective of the LCG is to represent prerequisite relationships in a concept space framed by the selected KB. The prerequisite relationship $pre(c_a, c_b) \equiv [c_a \quad isPrerequisiteOf \quad c_b]$ meets the following properties:(1) asymmetric $pre(c_a, c_b) \Rightarrow \neg pre(c_b, c_a)$, (2) irreflexible $pre(c_a, c_a)$ is not defined, (3) transitivity $pre(c_a, c_b), pre(c_b, c_c) \Rightarrow pre(c_a, c_c)$. In order to measure the prerequisite relations among concepts, we propose an approach to infer candidate relationships exploiting the graph base structure and the semantic of the concepts in the KB. Our proposal to find candidate prerequisite relationships is inspired in RefD [11] measure that fulfills the properties of our prerequisite relationship and shows superior results than more sophisticated supervised approaches. RefD infers prerequisite dependencies analyzing observable reference relationships between concepts (a reference could be a hyperlink, a citation, a mention, etc). This measure is defined as:

$$RefD(c_a, c_b) = \frac{\sum_{i=1}^{k} r(c_i, c_b) w(c_i, c_a)}{\sum_{i=1}^{k} w(c_i, c_b)} - \frac{\sum_{i=1}^{k} r(c_i, c_a) w(c_i, c_b)}{\sum_{i=1}^{k} w(c_i, c_b)} \qquad (1)$$

Where k is the size of the concept universe, $r(c_i, c_a)$ is an indicator function showing the existence of a reference between c_i and c_a, and $w(c_i, c_a)$ weights

the importance of c_i to c_a. In order to leverage RefD with the rich semantic relationships present in LOD KB, we will explore the use of hierarchical information (i.e. categories and other taxonomical information) and common related concepts follow the ontology properties to calculate $w(c_i, c_a)$. Likewise, following the properties paths [1] between concepts is an indicative of reference. Hence, different ontology properties of the KB could be explored for $r(c_i, c_a)$.

Recently, different similarity measures for LOD-KB have been proposed [19, 16] and appear to be a promising option to calculate $w(c_i, c_a)$. In preliminary results, we employed the similarity measures used by [20, 19].

Automatic Learning Content Sequence (ALCS): The ALCS has two main purposes. First, it builds a concept based representation for learning resources (Definition 1). This representation allows to enrich resources with metadata about the main topic concepts that they address. Second, given an ordered sequence of concepts extracted from LCG, it selects a ordered sequence of learning resources to address these concepts.

Definition 1: A resource r_i is represented as $R_i = \{(c, w(r_i, c))|c \in C\}$. The weight that denotes how the concept c is covered by the resource is defined by $w(r_i, c)$. The concept space C is limited by the selected LOD-KB, so each concept is basically an URI.

(1) Concept Based Resource Representation: The advantage of the bag of concepts based approach (Definition 1) is that it allows to exploit semantic information found in the KB to leverage and extend the representation. In the first step, the system extracts KB concepts mentions found in the text. This "enrichment" process takes an input text and returns a set of URIs of structured LOD entities to allow further reasoning on related concepts. Services such as DBpedia Spotlight, OpenCalais, AlchemyAPI or Bioportal could be employed for this task. Then, on the top of these mentions, different semantic expansion strategies could be applied [13]: (i) *categorical expansion (CE)*: for each annotation, we include its categories (or other hierarchical information about concepts in the KB) in the representation. (ii) *property Expansion (PE):* for each annotation, the representation is enriched with the concepts found by traversing some of the properties of the KB ontology.

We have found that large documents may involve concepts that have no relation to the main topic addressed by the resource, thus the above expansion strategies could lead to additional noise. Hence, instead of using such strategy, we proposed a filtering strategy that has already shown better results in this scenario. The concept behind this filtering strategy is that noisy concepts tend to be disconnected from those that represent the main topic of the resource. The filtering is achieved analyzing how the concepts are related via the linkages in the KB and discarding those that are disconnected or are weakly connected [13].

For the weighted strategy (i.e. $w(r_i, c)$ in Definition 1) we have followed approaches such as the frequency of the concept and an abstraction of the well-known TF-IDF [21]. These weighting schemas has shown to be enough to represent the coverage of the concept for a given resource. Concepts with higher

[1] https://www.w3.org/TR/sparql11-property-paths/

weights and its prerequisites extracted from LCG are incorporated as main concepts and prerequisites concepts metadata information.

(2) Resource Selection: Given a learning concept target c_j (i.e. a learning goal), ALCS extracts an organized sequence of concepts from LCG $g = \{c_1, c_2, ..., c_j\}$ to achieve it. Then, we select the set of resources $lr = \{r_1, r_2, ..., r_j\}$ that maximizes the coverage of the concepts in g. We follow a greedy approach that selects for each concept c_k in the sequence the resource r_i with highest $w(r_i, c_k)$. We plan to explore more sophisticated resource selection strategies based on the definition of a constraint satisfaction problem. Likewise, we will also investigate how to limit the number of prerequisites retrieved ($|g|$) in a "intelligent" way.

6 Evaluation Plan and Preliminary Results

We plan to evaluate the prerequisite relationships in LCG using the CrowdComp dataset [24]. This dataset contains binary-labeled concept pairs from five different domains collected using the Amazon Mechanical Turk. We plan to reported the accuracy, precision, recall, and F1 as evaluation measures. We want to compare our approach to discover prerequisites with the strategies reported by [24, 11, 12] (Q1,Q2). Some trials on a limited set of concepts, the use of RefD and LOD similarities explain above allowed us to test our approach on a small scale.

In order to test the information extraction and the resource representation (Q3,Q4) we have built a dataset with more than 30,000 documents that includes open academic articles recovered from CORE service[2], Web pages with technology and science news and a set of learning objects retrieved from MERLOT[3]. The complete and some parts of the dataset has already been employed in [14, 13]. We also plan to use a recent release dataset [5] that contains all embedded Learning Resource Metadata Initiative (LRMI)[4] markup statements extracted from the Common Crawl releases 2013-2015. Each entity description contains the URL of the WEB document from which it has been extracted, so it is possible to access the WEB resource content and employed our learning resource representation.

So far, a developed service that maps documents to bag of concepts representations (Definition 1) with expansion and filtering strategies was developed. This service could operate with different LOD-KBs, though all our experiments have been performed using DBpedia. We have partially evaluated the learning resource representation in an academic document recommendation task (Q3,Q4) [13, 15]. All academic documents in the candidate recommendation set were represented as R_i and different semantic expansion and filtering strategies were applied. In terms of classical measures of Top-N recommender tasks: MRR (Mean Reciprocal Rank), MAP@N (Mean Average Precision), and NDCG@N (Normalized Discounted Cumulative Gain), we have obtained superior results that word-based representations. In general our results validates that: (i) the proposal concept based representation express correctly an academic document content

[2] https://core.ac.uk/
[3] https://www.merlot.org/
[4] http://lrmi.dublincore.net/

(ii) using a concept space framed by DBpedia was enough to achieve a correct representation, so it is expected that specific domain ontologies could lead to better results and (iii) different strategies that exploit the background knowledge present in the KB could be employed to leverage the representation. We plan to evaluate how well this representation lead to correct metadata (i.e. main concepts) with human reviewers that manual annotate the resources with its main concepts.

7 Reflections

This research addresses the problem of finding and organizing learning resources to support self-directed learners in achieving a learning goal. All the proposed strategies have been oriented to the use of semantic technologies and, in particular, to the use of LOD-KB. Given a need for learning, the system draws a sequence that expresses the concepts involved and how they should be addressed. The organization of the sequence can be driven by some property of the KB's ontology, or through prerequisite relationships between the concepts inferred automatically from the background knowledge in the KB. To ensure that the sequence of concepts is translated into a sequence of resources, we propose a representation based on concepts automatically extracted from these resources contents. This representation allows to enrich the resource with metadata about the main concepts it addresses. We have still addressing the problem of generating the LCG that expresses the prerequisite relationships amongst concepts.

Acknowledgment. This thesis is funded by Colciencias under PhD scholarship No. 647/2014. I would like to express my heartfelt gratitude to my supervisor Dr. Olga. Mariño for her continuous encouragement. Likewise, I would also like to thank Prof. Stefan Dietze and the anonymous reviewers for their feedback.

References

1. Acampora, G., Loia, V., Gaeta, M.: Exploring e-learning knowledge through ontological memetic agents. IEEE Computational Intelligence Magazine 5(2), 66–77 (2010)
2. Al-Muhaideb, S., Menai, M.E.B.: Evolutionary computation approaches to the curriculum sequencing problem. Natural Computing 10(2), 891–920 (2011)
3. Bariso, E.U.: Personalised eLearning in Further Education. Technology-Supported Environments for Personalized Learning (2010)
4. Changuel, S., Labroche, N., Bouchon-Meunier, B.: Resources sequencing using automatic prerequisite–outcome annotation. ACM Trans. Intell. Syst. Technol. 6(1), 6:1–6:30 (Mar 2015)
5. Dietze, S., Taibi, D., Yu, R., Barker, P., d'Aquin, M.: Analysing and improving embedded markup of learning resources on the web. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 283–292. WWW '17 Companion, Republic and Canton of Geneva, Switzerland (2017)
6. Farhat, R., Jebali, B., Jemni, M.: Ontology based semantic metadata extraction system for learning objects, pp. 247–250 (2015)
7. Garrido, A., Morales, L., Serina, I.: On the use of case-based planning for e-learning personalization. Expert Systems with Applications 60, 1–15 (2016)

8. Hill, J.R.: Resource-Based Learning, pp. 2850–2852. Springer US, Boston, MA (2012)
9. Jebali, B., Farhat, R.: Ontology-based semantic metadata extraction approach. In: 2013 International Conference on Electrical Engineering and Software Applications. pp. 1–5 (March 2013)
10. Krieger, K., Schneider, J., Nywelt, C., Rösner, D.: Creating semantic fingerprints for web documents. In: Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics. pp. 11:1–11:6. WIMS '15 (2015)
11. Liang, C., Wu, Z., Huang, W., Giles, C.L.: Measuring prerequisite relations among concepts. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1668–1674. Lisbon, Portugal (September 2015)
12. Liang, C., Ye, J., Wu, Z., Pursel, B., Giles, C.L.: Recovering concept prerequisite relations from university course dependencies. In: In the 7th Symposium on Educational Advances in Artificial Intelligence (2017)
13. Manrique, R., Herazo, O., Marino, O.: Exploring the use of linked open data for user research interest modeling. In: 12mo Computing Colombian Conference (12CCC) (2017)
14. Manrique, R., Marino, O.: Diversified semantic query reformulation. In: The International Conference on Knowledge Engineering and Semantic Web (KESW2017) (Submitted) (2017)
15. Manrique, R., Marino, O.: How does the size of a document affect linked open data user modeling strategies. In: Int. Workshop on Web Personalization, Recommender Systems and Social Media (WPRSM 2017) (2017)
16. Meymandpour, R., Davis, J.G.: A semantic similarity measure for linked data: An information content-based approach. Knowledge-Based Systems 109, 276 – 293 (2016)
17. O'Donnell, E., Lawless, S., Sharp, M., Wade, V.P.: A Review of Personalised E-Learning: towards supporting learner diversity. International Journal of Distance Education Technologies 13(1), 22–47 (2015)
18. Paquette, G., Mariño, O., Rogozan, D., Léonard, M.: Competency-based personalization for massive online learning. Smart Learning Environments 2(1), 4 (2015)
19. Paul, C., Rettinger, A., Mogadala, A., Knoblock, C.A., Szekely, P.: Efficient Graph-Based Document Similarity, pp. 334–349. Cham (2016)
20. Pekar, V., Staab, S.: Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision. Proceedings of the 19th international conference on Computational linguistics - Volume 1 pp. 1–7 (2002)
21. Piao, G., Breslin, J.G.: Analyzing aggregated semantics-enabled user modeling on google+ and twitter for personalized link recommendations. In: Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization. pp. 105–109. UMAP '16 (2016)
22. Roy, D., Sarkar, S., Ghose, S.: Automatic extraction of pedagogic metadata from learning content. Int. J. Artif. Intell. Ed. 18(2), 97–118 (Apr 2008)
23. Sharma, R., Banati, H., Bedi, P.: Adaptive Content Sequencing for e-Learning Courses Using Ant Colony Optimization, pp. 579–590 (2012)
24. Talukdar, P.P., Cohen, W.W.: Crowdsourced comprehension: Predicting prerequisite structure in wikipedia. In: Proceedings of the Seventh Workshop on Building Educational Applications Using NLP. Stroudsburg, PA, USA (2012)
25. Wang, S., Ororbia, A., Wu, Z., Williams, K., Liang, C., Pursel, B., Giles, C.L.: Using prerequisites to extract concept maps from textbooks. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. pp. 317–326. CIKM '16 (2016)