Discovering Educational Resources on the Web for Technology Enhanced Learning Applications

Matteo Lombardi

School of Information and Communication Technology,
Griffith University,
170 Kessels Road, Nathan, QLD, 4111 Australia
matteo.lombardi@griffithuni.edu.au

Abstract. There are many works in TEL literature aiming to support learners and educators in their educational tasks. A current trend in this field is to develop expert systems and agents able to retrieve the most appropriate educational material. Such systems are great when handling pools of resources specifically designed for education, where the material is associated to educational data about the characteristics of the resource itself and its context of usage. Instead, many Web resources potentially suitable for education are not considered in their reasoning, because of the lack of such educational information. This PhD project aims to propose a solution for exploiting the Internet as an enormous repository of educational material, through an efficient discovering of Web resources that are suitable for education. Then, educational features of the resources shall be deduced and associated to them, building an overall concept map depicting the knowledge about educational material available on the Web. Current and novel TEL technologies are expected to benefit from the proposed research that will provide to them i) a novel huge amount of educational resources, and ii) a reusable representation of the educational knowledge coming from educators and institutions from all over the world.

Keywords: Web Crawling, Technology Enhanced Learning, Educational Resources, Concept Map

1 Introduction and motivations

The Web is widely recognized as a source of information for many different applications, including educational ones. There are many websites and online platforms which offer educational material covering a huge range of topics. In this scenario, learners and educators have the chance to access a large number of Web resources anywhere and anytime, with many benefits for them. Unsurprisingly, the Internet is now one of the most popular sources of information for the retrieval of teaching resources [16]. Another benefit for learners and educators is the increasing number of online systems designed for sharing educational resources. A current trend in education is offering Massive Open Online Courses (MOOC), namely online courses open to users from all over the world. Coursera[1], developed by Stanford University, is one of the most popular

[1] https://www.coursera.org/

MOOC platforms [19]. At the moment, Coursera hosts more than 2,000 courses from around 140 universities. The Internet is also plenty of platforms for sharing resources. Among the most popular there are SlideShare[2], YouTube[3], and Wikipedia[4].

The increasing trend of sharing educational resources on the Web has attracted several contributions from the research community. In fact, the research community in Technology Enhanced Learning (TEL) has produced contributions about the use of technology for the improvement of both learning and teaching processes [16]. Many of those works are proposals on the recommendation of educational resources [16, 27, 7, 23]. Since a substantial part of TEL contributions are towards the recommendation of Learning Objects (LO) [28] hosted on online Learning Object Repositories (LOR), it is clear that experts in TEL consider the Internet as a source of educational material. Therefore, it could be interesting to use the entire Web as a repository of teaching material.

However, the retrieval of Web resources presents some critical issues, mostly due to the fact that the Internet is an enormous and unorganized space. Search engines like Google are able to extract thousands of Web pages exploiting a user query, but it is not possible to filter the results according to a desired purpose (e.g. extract only material suitable for education). Even the most acknowledged proposals for crawling the Web do not assure to retrieve Web pages appropriate for specific purposes such as teaching. Indeed, approaches like the focused crawling [8] and the semantic crawling [17] are remarkable in extracting Web resources according to a set of topics or terms of interests, but the purpose of the resources is not taken into account.

There are several contributions that aim to structure the content of the Web, in particular for TEL applications [1, 21, 31]. Regarding to that, one of the most popular approach is to provide additional information to Web resources using metadata for LO. A number of standards have been proposed for annotating LO metadata, such as the IEEE Learning Object Metadata (LOM)[5], Dublin Core (DC)[6], and ADL SCORM[7]. Nevertheless, annotating metadata is an additional task that has to be performed mostly manually by LO authors. Such extra effort required is one of the drawbacks of composing, sharing and reuse LOs [26]. Recently, Linked Data (LD) [13, 4] is emerged has the new standard for contextualizing learning material [30, 32, 20, 10, 9], with the aim to improve [14] the previously discussed LO. However, other LO issues such as the diversity of metadata standards and their lack in representing some educational aspects are still present also in LD.

The first object of this doctoral project is to design and propose a new approach for Web crawling tailored to the educational field, able to fulfil the aforementioned gap performing an accurate extraction of resources suitable for teaching and learning, without any restriction on topics or terms. Then, the sharing and reuse of educational Web resources is expected to be promoted proposing an automatic extraction of edu-

[2] http://www.slideshare.net/

[3] http://www.youtube.com/

[4] http://www.wikipedia.org/

[5] IEEE 1484.12.1-2002, IEEE standard for Learning Object Metadata

[6] http://dublincore.org/documents/dces/

[7] http://www.adlnet.gov/scorm/scorm-2004-4th/

cational features, in order to represent two groups of data: the characteristics proper of the resources, and the context where the resource has been used (including other related resources). Also in [12], the educational features of a resource are divided in the same two groups.

Once the resources are equipped with contextual data, this PhD is expected to be concluded proposing a representation of the overall knowledge about educational Web resources in a Concept Map [6], a popular solution in TEL literature [22, 16]. Those findings are expected to be beneficial to researchers in Technology Enhanced Learning and Information Retrieval, towards a more effective support to students and educators in their educational tasks.

2 Related works

The first step of this research regards the identification and extraction of Web resources that are potentially useful for educational usages. Some important contributions in literature presented in this section are about the definition of educational characteristics of resources. Learning Object (LO) metadata standards are widely recognized by the research community as correct ways for representing educational information about a resource. Therefore, this study exploits the popular LO metadata standards IEEE Learning Object Metadata (LOM) and Dublin Core (DC), together with related works with the goal of discovering which educational characteristics are important when describing an educational resource.

Many works in TEL literature are focused on the feature selection and extraction processes. An interesting contribution in this scope [21] is the proposal of an automatic building of LOs using unstructured Web resources manually filtered by humans. In that work, the author selected a subset of LOM features that are important to be deduced for describing the educational characteristics of the resource itself. Another related work is the proposal of a framework for crawling Web resources and extracting their educational metadata [1]. In that study, focused crawling is used for restricting the mining to a domain deduced from a query given in input, and then metadata extraction is performed. The focused crawling approach can be used only when there are topics in input, so it cannot be applied to my research. About the extraction of educational metadata from Web resources, the authors suggest to represent the resources as vectors of terms, following a Vector Space Modeling (VSM) representation. Each term is then weighted according to its significance for the topic, so that the similarity among Web pages can be computed. A framework for analyzing textual resources gathered from the English version of Wikipedia is presented in [31]. In such work, the most important pieces of information are i) the article name, and ii) the text of the token used by Wikipedia (namely, the exact text used in the article for referring to another page).

A number of interesting approaches for developing Web crawling algorithms have been presented by the research community. Among them, the focused crawling approach [8] is defined as a selective seeking of Web pages that are relevant to a predefined set of topics, deduced from the analysis of documents [8] or Web pages [2] selected by the user, or from a set of terms [3]. Another suggestion is to estimate the relevance of a Web page before to visit it [25]. An interesting refinement to this ap-

proach is the Self-Adaptive Semantic Focused (SASF) crawler [15], where a module for learning patterns of pages of interest is involved for improving the filtering of novel Web pages. Although this approach is promising, SASF does not show a substantial improvement when compared to other Web crawlers. Another popular crawling approach is the semantic crawling [17], that aims to discover Web pages exploiting an ontology of terms connected by semantic relationships. Such ontology represents the knowledge of interest and it can be defined directly by users or using textual documents. A recent contribution applies such approach for discovering Web resources about environment and forecasting [29], using topic directories such as the Open Directory Project[8] for retrieving a set of words of interest for the specific domain.

Although Web crawling is a very popular topic, specially among researchers in Information Retrieval, there is still a gap in current approaches because they are tailored on topics, terms and domains, but not on the context of usage of Web resources. Therefore, the new crawling approach proposed in this work is expected to fulfil the current gap in the state-of-the-art and to unveil currently unclassified Web resources for education, overcoming the limit of topic and domain specificity.

The Web itself contains many known sources of educational material with semantic information. For instance, there are repositories of LOs such as MERLOT[9] [5], Connexions[10] and ARIADNE[11] that are very popular among TEL users and researchers. Several proposals for improving the retrieval of LOs [23] and for comparing the performance of systems based on them [24] have been presented. Open Educational Resources (OER) are another example of resources enriched with semantic data, in this case LD. An example of an institution that uses OER is presented in [10]. In that paper, the author depicts the Open University's Linked Data platform[12], an open-access system that aims to expose the public information of such university through LD. Among other information, learning materials described as OER are shared. This is now very common among universities and institutions, and there are even common platforms where OER are made publicly available[13]. LOs and OER sources are expected to contain an important number of educational resources. Hence, they will be explored during the crawling process, exploiting the associated metadata and LD for the successive feature extraction.

3 Research questions

This doctoral project is articulated in three main steps, each of them with a research question to be fulfilled. The first one is the following:

RQ1: What are the features that describe an educational resource?

[8] http://www.dmoz.org/
[9] http://www.merlot.org/
[10] http://cnx.org/
[11] http://www.ariadne-eu.org/
[12] http://data.open.ac.uk
[13] https://www.oercommons.org/

The review of popular metadata standards for annotating LO and the literature coming from the TEL research community has been fundamental for defining which features a resource suitable for education should have. Some of them like title and URL can be extracted from any Web resource, whether or not appropriate for education. Resources coming from MOOCs are among the only ones for which their educational features can be easily deduced, because such platforms present a standard structure that facilitates the feature extraction. Other sources of information that offer contextual data of resources are the repositories of LOs and OER already mentioned in the previous section of this contribution. So far, this research has exploited metadata standards, linked data dictionaries and the works presented in the literature review for proposing a comprehensive set of features, which are the most popular features used by educational platforms for describing their resources. Such set of features is the answer to this research question. Unfortunately, the majority of the remaining Web sites on the Internet are not structured for offering educational data. Then, deducing features very specific for education (like prerequisites, educational level and difficulty) from resources coming from other platforms than MOOCs, LOs and OER repositories is a foreseen challenge for the research activity.

Nevertheless, my proposal aims to determine if a resource coming from the Web is appropriate to be used as educational resource. Therefore, the second main research question to be addresses during the PhD program is:

RQ2: How to perform an efficient extraction of Web resources suitable for education?

For addressing the previously mentioned gap in literature about Web crawling, this research proposes to exploit existent MOOCs, LORs and OER platforms as sources of i) a huge amount of resources, and ii) precious teaching knowledge. In this scope, the PhD includes the development of a classifier able to compare resources coming from the Web with others hosted on the aforementioned educational platforms, namely trustworthy educational material. The proposed system shall classify a resource coming from the Web either as *suitable for education* or as *not appropriate for education*, considering its similarity to other resources already used in an educational context (i.e., already classified as *suitable for education*). At this stage, only MOOCs, LOs an OER repositories are considered because they host resources developed by human instructors and used in real courses. The choice of considering only such platforms in this phase allows to build a highly reliable set of educational resources in a short time. Up until now, this research has extracted more than 20,000 resources from Coursera creating a dataset of educational resources called DAJEE (DAtaset of Joint Educational Entities) [18], all of them described by the list of features discovered during the first step of this research. Table 1 shows the entities currently included in DAJEE. It is necessary to specify that the crawling of Coursera has been performed in August 2015, when preview of the course content was still possible. Moreover, the data has been extracted without infringing the copyright of the content, fully respecting the terms and conditions of Coursera.

Once the classifier is enough accurate in recognizing educational Web resources, an even bigger number of educational resources coming from other Web sites like Wikipedia, SlideShare and YouTube are expected to be included in the DAJEE dataset. Then, such new entries will be involved in the classification process as trusted educational resources. In this way, the system is expected to learn many other structures

Table 1. Summary of educational entities of Coursera included in DAJEE.

Entity	Number of crawled instances
University	99
Instructor	484
Course	407
Lesson	2,365
Concept	8,716
Resource	22,663
Transcript	14,327 (video resources only)

of educational resources and to improve the discovery of Web resources suitable for teaching and learning.

For concluding the PhD project, the third and last main research question to be satisfied is:

RQ3: How to represent the discovered knowledge in a data structure compatible with current technologies in TEL?

As anticipated in the previous section, building a Concept Map is one of the most popular ways for representing knowledge in education. Many proposals in TEL literature are about expert systems and agents using concept maps for their reasoning [23, 16]. Thus, the proposal of an overall concept map of the Web resources discovered as appropriate to be used in education is expected to be beneficial for already existent and novel researches. However, a concept map is useful when the elements are connected by some kind of semantic relationships. One of the last challenges of this research is the definition of educational relationships among the resources, starting from the structure of the MOOCs where resources are delivered, and the semantic information found on LOs and OER repositories [11]. Then, the relationships discovered so far shall be exploited for discovering connections also among novel Web resources classified as educational.

4 Potential applications

The overall goal of this PhD project is to propose a solution for i) extracting educational resources from the Web, and ii) representing the discovered knowledge as a concept map where resources are entities connected by educational relationships. Despite the number of interesting proposals in TEL for supporting the retrieval of learning material, at the moment, generic search engines like Google are still preferred by the users when looking for resources that are suitable for their educational tasks [5]. In this scope, the contribution of this doctoral project would be a very important step towards offering better educational applications both to students and to educators. Researchers in TEL are expected to benefit from a crawler tailored to education, which could be the first component of a novel educational-oriented search engine.

In order to achieve the overall goal of this research, the definition of the features that characterize educational resources is fundamental. Among many proposals and

standards for the representation of educational traits of digital material, this research provides the definition of a comprehensive structure compatible with current systems and useful for future contributions in TEL.

Other applications may exploit the overall concept map for helping educators in delivering their courses. Indeed, the educational knowledge contained into it has a huge potential that can support the building of a course from scratch, e.g., suggesting the most popular concepts and resources for teaching a subject. Even the refinement of already existent courses would be easier, exploiting the semantic relationships among resources used by other colleagues in their courses. Novel and up-to-date resources may be added to a current pool after a certain amount of time, or at the beginning of the academic term. In this way, this research shall also foster collaborations among institutions and educators from all over the world.

References

1. Atkinson, J., Gonzalez, A., Munoz, M., Astudillo, H.: Web metadata extraction and semantic indexing for learning objects extraction. Applied Intelligence 41(2), 649–664 (2014)
2. Batsakis, S., Petrakis, E.G., Milios, E.: Improving the performance of focused web crawlers. Data & Knowledge Engineering 68(10), 1001–1013 (2009)
3. Bedi, P., Thukral, A., Banati, H.: Focused crawling of tagged web resources using ontology. Computers & Electrical Engineering 39(2), 613–628 (2013)
4. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far (2009)
5. Brent, I., Gibbs, G.R., Gruszczynska, A.K.: Obstacles to creating and finding open educational resources: the case of research methods in the social sciences. Journal of Interactive Media in Education 2012(1), Art–5 (2012)
6. Cañas, A.J., Novak, J.D.: Concept mapping using cmap tools to enhance meaningful learning. In: Knowledge Cartography, pp. 25–46. Springer (2008)
7. Casali, A., Gerling, V., Deco, C., Bender, C.: A recommender system for learning objects personalized retrieval. Educational Recommender Systems and Technologies: Practices and Challenges, Hershey, PA: Information Science Reference pp. 182–210 (2012)
8. Chakrabarti, S., Van den Berg, M., Dom, B.: Focused crawling: a new approach to topic-specific web resource discovery. Computer Networks 31(11), 1623–1640 (1999)
9. d'Aquin, M.: Linked data for open and distance learning. Commonwealth of Learning Reports (2012)
10. d'Aquin, M.: Putting linked data to use in a large higher-education organisation. Interacting with Linked Data (ILD 2012) p. 9 (2012)
11. De Medio, C., Gasparetti, F., Limongelli, C., Lombardi, M., Marani, A., Sciarrone, F., Temperini, M.: Discovering prerequisite relationships among learning objects: a coursera-driven approach. In: Proceedings of the 15th International Conference on Web-Based Learning. Springer (2016)
12. De Medio, C., Gasparetti, F., Limongelli, C., Lombardi, M., Marani, A., Sciarrone, F., Temperini, M.: Towards a characterization of educational material: an analysis of coursera resources. In: Proceedings of the 1st International Symposium on Emerging Technologies for Education. Springer (2016)
13. Dietze, S., Sanchez-Alonso, S., Ebner, H., Qing Yu, H., Giordano, D., Marenzi, I., Pereira Nunes, B.: Interlinking educational resources and the web of data: A survey of challenges and approaches. Program 47(1), 60–91 (2013)

14. Dietze, S., Yu, H.Q., Giordano, D., Kaldoudi, E., Dovrolis, N., Taibi, D.: Linked education: interlinking educational resources and the web of data. In: Proceedings of the 27th Annual ACM Symposium on Applied Computing. pp. 366–371. ACM (2012)
15. Dong, H., Hussain, F.K.: Self-adaptive semantic focused crawler for mining services information discovery. Industrial Informatics, IEEE Transactions on 10(2), 1616–1626 (2014)
16. Drachsler, H., Verbert, K., Santos, O.C., Manouselis, N.: Panorama of recommender systems to support learning. In: Recommender systems handbook, pp. 421–451. Springer (2015)
17. Ehrig, M., Maedche, A.: Ontology-focused crawling of web documents. In: Proceedings of the 2003 ACM symposium on Applied computing. pp. 1174–1178. ACM (2003)
18. 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)
19. Kay, J., Reimann, P., Diebold, E., Kummerfeld, B.: Moocs: So many learners, so much potential. Technology 52(1), 49–67 (2013)
20. Keßler, C., d'Aquin, M., Dietze, S.: Linked data for science and education. Semantic Web 4(1), 1–2 (2013)
21. Krieger, K.: Creating learning material from web resources. In: The Semantic Web. Latest Advances and New Domains, pp. 721–730. Springer (2015)
22. Limongelli, C., Lombardi, M., Marani, A., Sciarrone, F.: A teacher model to speed up the process of building courses. In: Human-Computer Interaction. Applications and Services, pp. 434–443. Springer (2013)
23. 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 22(1–2), 58–82 (2015)
24. Lombardi, M., Marani, A.: A comparative framework to evaluate recommender systems in technology enhanced learning: a case study. In: Advances in Artificial Intelligence and Its Applications, pp. 155–170. Springer (2015)
25. Meusel, R., Mika, P., Blanco, R.: Focused crawling for structured data. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. pp. 1039–1048. ACM (2014)
26. Ochoa, X., Duval, E.: Quantitative analysis of learning object repositories. Learning Technologies, IEEE Transactions on 2(3), 226–238 (2009)
27. Sergis, S., Sampson, D.: Learning object recommendations for teachers based on elicited ict competence profiles. Learning Technologies, IEEE Transactions on (2015)
28. Sosteric, M., Hesemeier, S.: When is a learning object not an object: A first step towards a theory of learning objects. The International Review of Research in Open and Distance Learning 3(2) (2002)
29. Tsikrika, T., Moumtzidou, A., Vrochidis, S., Kompatsiaris, I.: Focussed crawling of environmental web resources based on the combination of multimedia evidence. Multimedia Tools and Applications pp. 1–25 (2015)
30. Vega-Gorgojo, G., Asensio-Pérez, J.I., Gómez-Sánchez, E., Bote-Lorenzo, M.L., Munoz-Cristobal, J.A., Ruiz-Calleja, A.: A review of linked data proposals in the learning domain. Journal of Universal Computer Science 21(2), 326–364 (2015)
31. Wojtinnek, P.R., Pulman, S., Völker, J.: Building semantic networks from plain text and wikipedia with application to semantic relatedness and noun compound paraphrasing. International Journal of Semantic Computing 6(01), 67–91 (2012)
32. Zablith, F.: Interconnecting and enriching higher education programs using linked data. In: Proceedings of the 24th International Conference on World Wide Web Companion. pp. 711–716. International World Wide Web Conferences Steering Committee (2015)