Creating Learning Material from Web Resources

Katrin Krieger(✉)

Faculty of Computer Science, Knowledge-based Systems
and Document Processing Research Group, Otto-von-Guericke-University
Magdeburg, Magdeburg, Germany
katrin.krieger@ovgu.de

Abstract. We observed that learners use general Web resources as learning material. In order to overcome problems such as distraction and abandonment of a given learning task, we want to integrate these Web resources into Web-based learning systems and make them available as learning material within the learning context. We present an approach to generating learning material from Web resources that extracts a semantic fingerprint for these resources, obtains educational objectives, and publishes the learning material as Linked Data.

1 Problem Statement

Technology-enhanced learning (TEL), especially Web-based learning, has become a fundamental part in education over the last decades. E-Learning platforms provide access to electronic learning material, accompany in-class lectures in blended learning scenarios or offer assessment facilities for formal and informal testing. Whole courses are held online, whether as qualification training, school education in sparsely populated areas or as courses dealing with special topics, letting remotely located experts teach students all over the world. TEL has torn down barriers in time and space, enabling students to learn where and whenever they want.

Our research focuses on a blended learning scenario. Students attend both lectures and tutorial classes. The provision of learning material such as slides and scripts and the assignment of homework are handled via a learning management system (LMS). The LMS allows students to upload their programming homework via a website; it also checks the completed assignments and gives immediate feedback through e-assessment functionality[1].

In in our setting, we teach undergraduate students, who form a heterogeneous group with respect to previous knowledge and skills. Therefore, it is essential that each learner gets support in terms of additional assistance and feedback.

When students solve their homework assignments, they have to apply the theoretical input from the lecture to practical problems. This is a scenario, where they have to focus on the given task, thus a more directed approach to problem solving – in contrast to exploratory or inquiry-based learning – is reasonable.

Experience has shown that students having difficulty in an e-assessment session not only rely on the learning materials provided for the course, but search

[1] c.f. eduComponents, http://wdok.cs.uni-magdeburg.de/educomponents.

F. Gandon et al. (Eds.): ESWC 2015, LNCS 9088, pp. 721–730, 2015.
DOI: 10.1007/978-3-319-18818-8_45

the Web for additional materials that might be helpful. This interrupts their e-assessment session and might lead to distraction and even abandonment of the assigned task [10].

In general, learners seem to use conventional Web resources[2] as learning material. Our idea is to provide learning material in e-learning contexts that has been integrated from such general Web resources. We will analyze, how general Web resources can be linked to e-learning environments and develop a method to automatically integrate appropriate Web resources into the recent learning context of the LMS.

Our strategy is to offer learners additional learning material, which they can access immediately instead of interrupting the e-assessment session and turn to a Web search engine. Web resources will be automatically integrated into the recent learning context and presented as additional learning material in a didactical meaningful way.

2 Hypotheses

Learners use Web resources as learning material in educational contexts. Hence, we propose the following **hypotheses**:

1. It is possible to judge (automatically) whether the content of a Web resource is relevant with respect to a learning context or not.
2. Web resources carry data that can be used to derive information describing educational and didactical characteristics.
3. If a Web resource can be automatically structured and augmented with metadata as presumed, it is possible to integrate this Web resource into learning contexts in technology enhanced learning (TEL) systems such as LMS.

3 Research Question

The hypotheses stated previously lead to the **central research question** for this dissertation project:

> How can Web resources be structured and enriched with metadata such that they can be linked to e-learning contexts and act as learning material?

4 Approach

The hypotheses stated in Sect. 2 identify three aspects that have to be taken into consideration when we want to automatically create learning material that can be integrated into Web-based LMS:

[2] Artifacts found on the Web - documents, slides, videos, audio files, etc. will be referred to as Web resources throughout this paper.

1. The Web resource has to match the learning context. That means that the learning context and the Web resource have to be semantically closely related. Helpful learning material uses the same terms and definitions as the learning context. Hence, we have to take care that we generate a sufficiently accurate description of the content of the Web resource. This description can then be used to judge whether the resource is a candidate to act as learning material in a particular context.
2. The Web resource has to be augmented with educational metadata such that it can be integrated into the LMS in a didactical meaningful way. This educational metadata is about different pedagogical dimensions that can be used to filter and sort the learning material, enabling the learner to chose a material that might be the most suitable for his personal needs.
3. We want to close the gap between the e-learning environment and the Web. We want to help the learner to stay focused on his task and offer additional learning material that can be immediately accessed. This means that we need a seamless integration of the created learning material into our Web-based LMS. The de-facto standard for data integration on the Web is Linked Data. Therefore we will make the learning material with its semantic and educational description available as Linked Data - as so-called *Linked Learning Items (LLI).*

In the following subsections we will go into detail about how these aspects will be addressed.

4.1 Overall Process

In order to create learning material from Web resources, we have to define a process that takes a Web document as input and produces a Linked Learning Item as output. This process contains three stages that correspond with the aspects identified above.

We designed a REST-based framework with interchangeable services which realize the three aspects: create a semantic description of the input Web resource, extract educational metadata, and deliver the resulting LLI.

4.2 Semantic Fingerprint

To create a semantic description of the content of a Web resource that helps judging whether this resource is relevant as learning material in certain learning contexts, we will generate a structure we call *semantic fingerprint.* This fingerprint will be a graph-like structure, containing ontological concepts as vertices and relations between these concepts as edges.

A semantic fingerprint is generated in an iterative process:

1. Keywords are extracted from the resource. This set of keywords $K = \{k_1, \dots, k_n\}$ is the input for the following steps.

2. Graph nodes are created by mapping each keyword, that has been extracted from the resource, to a number of concepts: $C = \bigcup_{i=1}^{n} C_i$ for $C_i = \{c_1, c_2, \ldots, c_{|C_i|}\}$

   Each keyword k is mapped to a set with a certain number of concepts c. This is done by querying DBpedia's SPARQL endpoint for concepts that match the given keyword.

   At the end of this step, the graph contains only nodes, but no edges: $SF = (C, \emptyset)$.

   This set of nodes consists of relevant as well as irrelevant nodes, because mapping keywords to concepts will return also concepts, that have the same or similar label but represent different concepts. These irrelevant concepts will be removed later from the graph.
3. To determine connections between the nodes C we try to find paths between those nodes. We expand every node and look for neighboring concepts C_e, i.e. we perform a breadth first search. This is done by querying DBpedia's SPARQL endpoint for concepts that have a relation with the given concept.

   Further connections can be found with reasoning over the graph to reveal implicit relationships. The feasibility of applying other approaches like OWL API or the OWL entailment regime has to be investigated.

   Furthermore we will analyze how semantic relationships that are included in the text of the Web resource, but not in the ontology, could be extracted and added to the graph (e.g., with approaches from natural language processing, such as latent semantic analysis).

   This step introduces more nodes to the graph as well as semantic relationships as edges. The result is a graph $SF = (C \cup C_e, E)$ containing all concepts C, that can be mapped to the resource's keywords, neighboring nodes C_e, as well as their relationships.
4. The graph is cleaned by removing irrelevant edges and concepts. The result of this step is a graph consisting of several connected subgraphs. Each subgraph contains concepts about different topics, that are not semantically related. To identify irrelevant relations we use a number of heuristics.
5. We identify all connected subgraphs in the graph.
6. The semantically most relevant subgraph is chosen as the semantic fingerprint for the resource $SF = (C', E')$.

4.3 Educational Metadata

To enable an appropriate didactical representation of the LLI we need to describe it with educational metadata.

We have inspected and analyzed existing standard vocabularies and efforts (c.f. Sect. 8.2). The selection of educational metadata vocabularies for this project depends on two factors: the usefulness for the description within the LLI, i.e. it is useful for filtering and sorting; and the availability of data that can be automatically extracted or generated.

We identified several fields from the educational categories of the *Learning Objects Metadata standard*[3] (LOM) that will serve as educational metadata fields for the LLIs:

- interactivity type
- learning resource type
- semantic density
- description

The *interactivity type* can have the values *active, expositive, and mixed* and describes the level of interaction a learner can have with the learning resource. We will derive the value for this field from the type of the Web resource. When we know, that the resource is a video, we can conclude that this is an "expositive" resource.

The *learning resource type* can have a number of predefined values: *exercise, simulation, questionnaire, diagram, figure, graph, index, slide, table, narrative text, exam, experiment, problem statement, self assessment, and lecture.* We also want to derive the value from the Web resource type as well as from its content format.

Semantic density describes the degree of conciseness of a learning resource. We will exploit the semantic fingerprint to obtain a value for this field which can have the values *very low, low, medium, high, very high.* The size and shape of the semantic fingerprint might give insights about the semantic density. How we can derive a value for semantic density from the semantic fingerprint has still to be investigated.

The *description* is an open text element. The value can be obtained from the resource's title or its content description. We have made some experiments with Web 2.0 portals such as StackOverflow[4]. These portals offer APIs which enable and easy access on such data programmatically.

4.4 Publication as Linked Data

For a seamless integration of the Web resource along with its semantical and educational description we will deliver the learning material as an LLI. An LLI is a data object conform to Linked Data standards that enable a Web-based integration of the data as well as the possibility to share the LLI with others, e.g., in . A serialization of the LLI as JSON-LD[5] is planned.

5 Evaluation Plan

5.1 Hypothesis #1

> "It is possible to judge (automatically) whether the content of a Web resource is relevant with respect to a learning context or not."

[3] http://ltsc.ieee.org.

[4] http://www.stackoverflow.com.

[5] http://json-ld.org/.

A semantic fingerprint has certain desired properties:

P1: Concepts in the fingerprint are distinct and unambiguous. That means, that the fingerprint should contain only concepts, that describe the resource content as clearly as possible. Concepts, that refer to homonyms or polysems of keywords do not belong to the fingerprint and would only add noise.
P2: Concepts in the semantic fingerprint are connected through relations. A semantic fingerprint is a completely connected graph. Concepts, that are semantically related, are connected through an edge.
P3: Resources, that have semantically similar contents will yield similar fingerprints. This means, that those fingerprints contain common concepts and relations or other particular substructures in the graph.
P4: A semantic fingerprint covers all essential concepts that belong to the resource. Thus, all keywords from the resource that belong to a certain topic or area should relate to at least one concept in the resulting fingerprint.

Since we explicitly add relationship edges to the fingerprint during the generation process and remove such nodes, that are not connected to the graph, property P2 is always met. To show that our approach generates semantic fingerprints, that carry the desired properties P1, P3 and P4, as well as to demonstrate the robustness of the method, we will conduct an evaluation.

The first stage is a quantitative analysis where we examine the influence of the keywords on the generated fingerprints. In the second stage of the evaluation we let human reviewers rate the quality of the semantic fingerprints. Some first findings are shown in Sect. 6.

5.2 Hypothesis #2

> "Web resources carry data that can be used to derive information describing educational and didactical characteristics."

The LLIs contain a description with educational metadata. This description includes fields for the different dimensions of such data, such as *interactivity type*, *difficulty*, or *learning resource type*. The approach to extract and generate the values for these elements can be considered successful when we find values for all elements. The quality of the collected educational metadata in terms of helpfulness and suitability in a learning context will be evaluated with human probands which will include learners as well as instructors. We plan to let at least 5 learners and 5 instructors rate the educational metadata of 20 LLIs with a questionnaire.

5.3 Hypothesis #3

> "If a Web resource can be automatically structured and augmented with meta- data as presumed, it is possible to integrate this Web resource into learning contexts in technology enhanced learning (TEL) systems such as LMS."

The LLIs are data objects that are compliant to the Linked Data principles. By adhering to these very principles throughout development and implementation we can make sure that the LLIs can be integrated into Web-based systems such as LMS. No formal evaluation is needed.

6 Preliminary Results

We have developed and implemented a REST-based Web service that will generate a Linked Learning Item from a Web resource.

An HTTP-based client can send a document or document context to the Web service. The Web service will compute the semantic fingerprint and educational metadata. It will return this data as an LLI. To generate the semantic fingerprint the Web service queries connectors to APIs of knowledge bases such as Freebase, DBpedia, or Wordnet, to match keywords from the resource to ontological concepts and discover relations between those concepts (see Sect. 4). Another component within the Web service will extract and generate values for the educational metadata. For testing purposes we implemented different connectors to Web 2.0 platforms such as Slideshare[6] and StackOverflow as well as a connector that indexes the content of lecture slides. We can query these connectors to get Web resources for which we then compute the LLI. Please note that these connectors do not belong to the core of the Web Service and have been developed only for the development of the LLI method.

Furthermore, we have developed an approach to create semantic fingerprints from Web resources and conducted an initial evaluation[7]. The evaluation revealed, that the size of the generated keyword list for a given document is crucial. Hence, the keyword extraction algorithm that will be used in conjunction with the fingerprint generation process should rather return more keywords than trying to prefilter them. Since the fingerprint generation process will eliminate irrelevant concepts it is not necessary to filter the keyword list. It should be preferred to create the fingerprint with a higher number of keywords.

7 Relevancy

The dissertation will contribute to several research fields since it concerns technology enhanced learning (TEL), Linked Data and the Semantic Web. We will inspect different methods and processes from these fields and analyze their usefulness for the solution of the stated problem. A combination of techniques will be deployed and tested to automatically augment Web resources with educationally relevant metadata and deliver an LLI.

The practical impact of this dissertation project will be as follows:

1. *Learners* will benefit from our new method while working with an LMS. They will be provided with new learning material that can be delivered instantaneously, e.g., during an e-assessment session. They can fully focus on solving

[6] http://slideshare.net.

[7] This work has been submitted as a conference article to COMPSAC 2015.

problems, because they will be supported with additional learning material within the system. They do not need to interrupt their session to consult a search engine for further information on the subject.

2. *Instructors* will benefit from a decreased workload for the creation of electronic learning material since additional learning material can be automatically created. The new method could be used in an authoring tool as a kind of recommendation service. During the creation of a course or assignment, the authoring tool could automatically fetch Linked Learning Items that match the recent learning context and offer them for integration as additional learning material.
3. *The Linked Data community*, especially the Linked Education community[8], will benefit from an important real-world application that turns legacy Web data into Linked Data-compliant data that can be integrated into Web-based e-learning environments.

8 Related Work

8.1 Semantic Fingerprints – Building Structured Data

With the Semantic Web and Linked Open Data developing quickly, there are various approaches to automatically map entities from unstructured text to LD entities and detect relations between those entities.

LODifier [1] uses Named Entity Recognition (NER) and maps the named entities to DBpedia URIs. However, in contrast to our approach, the detection of relations between such entities is done by means of statistical parsers and discourse representation structures. To disambiguate concept mappings, the authors use Wordnet mapping tools. The resulting structure is converted into an output format which is conform to RDF standards. References [2,7] follow a similar approach, but do not work on completely unstructured text. These methods are based on partially labeled data and therefore demand manual annotation as a preprocessing step, which is not necessary in our method. Reference [11] describes an approach to build semantic networks from plain text and Wikipedia pages, that relies only on linguistic tools and uses no other structured data as resource.

Reference [6] describes another approach to build ontologies from natural language texts by combining Discourse Representation Theory, linguistic frame semantics, and ontology design patterns.

We will look into combining our purely Linked Data-driven approach with linguistic tools used in [1,6,11] in order to create a more accurate semantic representation as the semantic fingerprint.

Web documents, or HTML documents in particular, carry a structure, that can be exploited for generating structured data.

Rowe describes in [8] a method for turning legacy Web pages about scientists and faculty staff into Linked Data utilizing the document object model (DOM) and Hidden Markov Models to build RDF triples and link those facts to the Web of Data.

[8] http://linkededucation.wordpress.com/.

SparqPlug also exploits the DOM of a legacy webpage to build Linked Data models. SPARQL queries are executed over an RDF model that has been created from the DOM in order to extract relevant data [3].

So far, we have not taken the exploitation of the HTML structure into account, but approaches like SparqPlug could be included into our precess to prestructure Web documents. This preprocessing step might be valuable when we extract keywords or concepts, respectively, that serve as input for the semantic fingerprint creation process.

8.2 Educational Metadata

There have been efforts to structure and formalize electronic learning material as so called Learning Objects (LO). These are pieces of digital learning content along with metadata containing a structured semantic description. The LOs reside in Learning Object repositories where users can lookup, re-use, and share those objects. The idea – in analogy to software components – was to build libraries with reusable learning material that can be assembled into a new LO, e.g. a course about computer graphics containing pieces of content from mathematics, physics, and computer science. This was meant to ease the re-use of already existent electronic learning material and the creation of new material as a mashup from other learning resources.

In order to create Web-based learning material that is ontologically annotated, there is a need for standard ontologies that cover different aspects of teaching and learning. Besides standard domain ontologies dealing with knowledge about particular domains, ontologies about pedagogical knowledge, e.g., curriculum sequencing, student modelling, grading and other pedagogical issues are required [4].

PASER [5] is a system for automatically synthesizing curricula for online courses. AI planning and Semantic Web technologies are used to combine appropriate learning objects into personalized online lectures. This approach includes different metadata standards such as LOM, content packing, educational objectives and learner related information. PASER employs an ontology for maintaining a hierarchy of competencies.

SIEG [9] is a system that creates learning objects for a certain e-learning application automatically from ontologies. These domain ontologies are deployed for different courses, such as WordNet is used to create learning objects about english grammar, or YAGO is used for learning objects about history. The resulting learning objects store triples, such that they can be used for learning facts.

Both approaches build on already existing educational metadata which is used to create the respective learning object. To our knowledge, there is no work on extracting and deriving educational metadata from unannotated documents yet.

Acknowledgements. I would like to express my gratitude to my Ph.D. advisor Prof. Dr. Dietmar Rösner for continuous support and encouragement. Likewise, I want to thank Prof. Sebastian Rudolph and the anonymous reviewers for their time and substantial feedback.

References

1. Augenstein, I., Padó, S., Rudolph, S.: LODifier: generating linked data from unstructured text. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) ESWC 2012. LNCS, vol. 7295, pp. 210–224. Springer, Heidelberg (2012)
2. Byrne, K., Klein, E.: Automatic extraction of archaeological events from text. In: Proceedings of Computer Applications and Quantitative Methods in Archaeology, pp. 1–16 (2010)
3. Coetzee, P., Heath, T., Motta, E.: Sparqplug: generating linked data from legacy HTML, SPARQL and the DOM. In: Bizer, C., Heath, T., Idehen, K., Berners-Lee, T. (eds.) Proceedings of the WWW 2008 Workshop on Linked Data on the Web. CEUR Workshop Proceedings, vol. 369. CEUR-WS.org (2008)
4. Devedžić, V.: Education and the semantic web. Int. J. Artif. Intell. Educ. **14**, 39–65 (2004)
5. Kontopoulos, E., Vrakas, D., Kokkoras, F., Bassiliades, N., Vlahavas, I.: An ontology-based planning system for e-course generation. Expert Syst. Appl. **35**(1), 398–406 (2008)
6. Presutti, V., Draicchio, F., Gangemi, A.: Knowledge extraction based on discourse representation theory and linguistic frames. In: ten Teije, A., et al. (eds.) EKAW 2012. LNCS, vol. 7603, pp. 114–129. Springer, Heidelberg (2012)
7. Ramakrishnan, C., Kochut, K.J., Sheth, A.P.: A framework for schema-driven relationship discovery from unstructured text. In: Cruz, I., et al. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 583–596. Springer, Heidelberg (2006)
8. Rowe, M.: Data.dcs: converting legacy data into linked data. In: Bizer, C., Heath, T., Berners-Lee, T., Hausenblas, M. (eds.) Proceedings of the WWW 2010 Workshop on Linked Data on the Web (LDOW 2010), vol. 628, April 2010
9. Soto, A., Hernández, J.A.F., de los Angeles Buenabad Arias, M, Diez, G.: Using ontologies to generate learning objects automatically. In: Gelbukh, A., Mendoza, M.G., Alcántara, O.H. (eds.) Proceedings of the 1st Workshop on Intelligent Learning Environments WILE 2009 (2009)
10. Winter, J., Cotton, D., Gavin, J., Yorke, J.: Effective e-learning? Multi-tasking, distractions and boundary management by graduate students in an online environment. Res. Learn. Technol. J. Assoc. Learn. Technol. (ALT) **18**(1), 71–83 (2010)
11. Wojtinnek, P.R., Völker, J., Pulman, S.: Building semantic networks from plain text and wikipedia with application to semantic relatedness and noun compound paraphrasing. Int. J. Seman. Comput. (IJSC) **6**(1), 67–91 (2012). Special Issue on Semantic Knowledge Representation