Semi-automated Semantic Annotation of Learning Resources by Identifying Layout Features

Sylvain DEHORS*, Catherine FARON-ZUCKER**, Jean-Paul STROMBONI**,
Alain GIBOIN*
**ACACIA, INRIA Sophia Antipolis*
2004 route des Lucioles, 06902 Sophia Antipolis cedex, France
Sylvain.Dehors,alain.Giboin@sophia.inria.fr
***MAINLINE, I3S, UNSA*
930 route des Colles, bât ESSI,06930 Sophia Antipolis cedex, France
faron,stromboni@essi.fr

Abstract. It is now widely accepted that any kind of digital content must be somehow semantically annotated to be intelligently used by computer programs. Annotations can be metadata, descriptions, etc. When dealing with learning, most systems require the author to manually annotate resources so that the system can deploy a navigation strategy, an adaptive behavior etc. However this task is very problematic, and often reveals to be an overwhelming enterprise. In this paper we propose a methodology, based on the reuse of existing pedagogical documents to achieve a semi-automated extraction of semantic annotations by identifying semantic information contained in the layout. We are applying this methodology in the design of a Web Based Learning System.

1 Introduction

The now classical approach to learning systems design is to rely on semantic annotations of pedagogical documents; this applies whether in the domain of LMS (Learning Management Systems), LCMS (Learning Content Management Systems) or LOR (Learning Object Repository). Exchanging metadata that are both understandable by humans and interpretable by machines is also the vision of the Semantic Web and languages, like RDF and OWL, are the key to express semantic annotations in a standard way.

In order to be used by computer systems, annotations must be expressed using a strict vocabulary often related to a model. This model, whether a "thesaurus", a "domain model" [1] or an "ontology" [2], provides common references to annotate resources.

One of the major pitfall, is the creation of those annotations manually (by humans), automatically (through automated programs), or semi-automatically (both by humans and programs). In this paper we first have a look at the existing approaches to annotate learning resources. Then we propose a methodology for semi-automatically extracting semantic annotations from pedagogical documents by identifying the semantic information contained in the layout. Finally we present the application of this method to the design, implementation and use of a simple WBLS based on semantic technology.

2 Existing methods and tools for the annotation of learning resources

As most learning systems use tailored courses, they require teachers to specifically create each document used by the system. Teachers are provided with authoring tools [1] to create new documents. Most of the existing research tools generate information in proprietary formats, whereas international standards like SCORM are being more and more enforced by commercial products. But this requires a lot of work and imposes major constraints upon the author. Moreover it does not take advantage of the huge amount of learning resources already available, both on the web or in the author's personal resources.

Another approach is then to consider reusing existing material by complementing it with extra annotations. To create such annotations, teachers are provided with annotation editors dedicated to learning resources. The interface appears often as a form to fill in and it is quite difficult to fill them with relevant and coherent information. For example several experts might not agree upon a document's content, and what concepts can be used to annotate it. Reusing existing material also implies to work on the size of the content; sometimes a resource can be a complete book [3]. To propose enhanced navigation it is necessary to slice the content into smaller chunks.

Here we propose a methodology for reusing document content and displaying it in a WBLS without relying on a specific annotation tool with form-based annotation.

3 Annotation of learning resources based on document layout features

The basic assumption we rely on is that the annotation task must be straightforward for the teachers and must not impose them to use the underlying formalism chosen for storing the annotation. For example it is not a viable option to manually edit HTML or RDF files.

We argue that every course is based on a learning or pedagogical model, which includes some pedagogical strategy. So first, the teacher is asked to explicit the pedagogical strategy for his/her course. For example in our case study we focused on a question-based approach to motivate learners to read the course. Then, the annotation task consists in interviewing the teacher, who is also the author of the document, and making him/her explicit the model of the existing document and how this model supports the envisioned educational strategy. Once this model is defined, the annotation task consists in identifying in the layout of the document the markers of the elements of this model. For example if the model defines the concept of "important notion of the domain", it is very likely that the corresponding word will appear in bold somewhere in the page, and so on. This kind of "visual" information must be gathered and standardized through a discussion with the author. Then a phase of re-authoring according to this layout principles must take place to ensure that all the visual clues are present to identify each component's role according to the model. The final step consists in formally creating instances of the components of the document through an automated process. Here we argue that the automation task is not very hard as most formats used for courses today (.doc, .ppt, .tex) are well structured. For example Word documents can be exported to XHTML and then treated by XSL transformation to extract annotations in the desired formalism (XML, RDF). We have successfully applied this methodology in the experiment described below.

4 A Web-based Learning System Design Experiment

We have developed a system called QBLS (for Question-Based Learning System) to demonstrate the use of semantic web technologies for setting and running a Web based learning system. The course we took for our first experiment is an introduction to signal analysis for first year computer science french engineering students. This course has previously given rise to thoughts about the use of information technologies in education. The aim of the experiment was to set up a knowledge pool where pedagogical resources are annotated in RDF and semantic queries are performed by the semantic search engine Corese [5].

4.1 Acquisition and representation of pedagogical knowledge

Our original document was a unique Microsoft PowerPoint File, supporting one hour of formal lecture. It was following an implicit model, for example the curriculum objectives were explicitly written at the top of every slide, and a set of relevant questions were given to motivate the students. Our task was then to formalize this underlying model that, we claim, exists in any pedagogical document of reasonable quality. This document was used as a support for oral teaching but was also given to the student as a hard copy course reference. This is a very common practice at university level, so this example is quite significant.

We had to model the document to reflect the pedagogical strategy chosen by the teacher. It is important to notice that our teacher was the author of the document and further investigation is needed to determine if someone else could have done it. This is crucial as reusing material is seen as the way to reduce the high-cost task of creating adaptable digital learning material. At least we demonstrated here that this was possible for the author to re-use his own documents. The defined model has no ambition of being generic, as our rationale is to save time and effort for teachers and the resources are to be used by a specifically designed WBLS.

4.2 Ontology of the pedagogical document

Because the final annotations would be expressed in RDF we formalized the model in an RDFS ontology. The pedagogical roles expressed in the model (like “definition”, “example”, etc.) formed the conceptual vocabulary we used to create the ontology. This ontology is a “pedagogical” one as defined in [2]. In the approach described in this paper the domain is solely represented by a set of concepts. It doesn’t seem to be possible to extract more information about the domain (like relationships between concepts) without a more sophisticated approach and then more work for the teacher. This approach is quite different from those relying on a model/ontology of the domain to “describe the content” of learning resources. Here the domain model comes solely from the document and only pedagogical information is to be used by the targeted WBLS. The ontology designed here is not meant for sharing resources across the globe but just between a teacher and his/her students, for more details see [6]. Once we had defined the ontology, a set of styles was created in Word. The teacher had to apply them on the document to identify the different components of the model. We only relied on the layout to generate annotations. For example words in italic, refer to the concepts in the course. It is important to notice that this was done with the sole use of the usual Microsoft Word program the teacher is familiar with. This technique was applied in [1] but we express much more information here.

4.3 Semantic-based retrieval of resources for visualization and navigation

QBLS uses the semantic search engine Corese [5] to perform queries on the RDF annotation base. We use it to retrieve the annotated resources on demand. Using the system the learner visualizes resources relative to a concept of the domain, and navigates from one to another resource either by following “seeAlso” links which leads to other concepts or by

gathering information on the same concept by accessing the definition, example(s), etc. available. According to the classification proposed by [1] our approach is to set up a “concept-based hyperspace”.
When asking for a new concept a query is sent to the Corese engine through a Web server that sends back an RDF result. That result is then displayed in a browser using an XSLT stylesheet. We must stress here that most of the development effort was spend on the generation of the annotations. The remaining part of presenting the resources in a web browser was done with very little development time, as most of the job is done by the existing search engine. The QBLS system has been used during a one hour formal course and a two hours exercise session by 49 students. The conceptual navigation provided by the system was quite appreciated by the students. They gave a high score (4.2 out of five) to a “system usability” questionnaire given after the session.

5 Conclusion

In this paper we have presented a methodology for semi-automatically extracting annotations from existing pedagogical documents. The corner stone of our method is that it does not require the use of any specific annotation tool but a little re-authoring work for the teacher, just manipulating the layout. For this re-authoring task the teacher can use the tool he is used to, which is a major incentive for him/her. This methodology also relies on the collaboration between the teacher and an ontologist/modeler to decide on the model of the document. This method relieves the teacher from the burden of authoring and heavily annotating the whole course. The application of our method to the design of a Web based leaning system has lead to an effective experiment showing that the semantic expressivity of the annotations acquired from the layout of the original document is quite sufficient to support a dynamic navigation in the so-built QBLS system. This work also pointed out some weaknesses of the existing standard models as they require far too much effort and maybe cannot even be effectively put in practice by a normal teacher. In the next experiment we will refine the methodology sketched out here and show more advantages of using standard semantic web technologies for designing and implementing WBLS.

References

[1] Brusilovsky, P. (2003) Developing adaptive educational hypermedia systems: From design models to authoring tools. In: T. Murray, S. Blessing and S. Ainsworth (eds.): Authoring Tools for Advanced Technology Learning Environment. Dordrecht: Kluwer Academic Publishers, 377-409.

[2] R. Mizoguchi, M. Ikeda, and K. Sinitsa (1997) Roles of Shared Ontology in AI-ED Research -- Intelligence, Conceptualization, Standardization, and Reusability. Presented at AIED-97, Kobe.

[3] P.Rigaux and N.Spyratos (2003) Metada Management and Learning Object Composition in a Self e-Learning Network, in proc. of Workshop on Information Search, Integration and Personalization, Japan.

[4] Stromboni J.P. (2002), Un cours introductif au traitement du signal à travers les applications audio de l’ordinateur multimédia, in Technologies de l’Information et de la Communication dans les Enseignements d’ingénieurs et dans l’industrie (TICE 2002), Lyon.

[5] Corby O., Dieng R. et Faron C. (2004), Querying the semantic web with the Corese search engine, in proc. of the European Conference on Artificial Intelligence (ECAI 2004), subconference PAIS, Valencia.

[6] Dehors S., Faron-Zucker C., Giboin A., Stromboni J.P. (2005), QBLS : web sémantique de formation pour un apprentissage par questionnement, In Environnements Informatiques pour l'Apprentissage Humain (EIAH 2005), Montpellier.