Automatic Extraction of Ontologies from Teaching Document Metadata

Judy Kay and Sam Holden
School of Information Technologies
University of Sydney, Australia

Abstract

SITS (Scrutable Intelligent Teaching System) is designed to make use of existing learning items in flexible and effective learning interactions. The reuse of pre-existing resources is important since creating new learning resources is a time consuming task which requires a skilled author. The Internet also provides a large number of resources for reuse. A major hurdle for reuse is metadata, especially epistemological metadata since different teachers or courses may have different ontologies for a given domain. SITS takes a minimalist approach to metadata. It requires only that the author of a teaching environment should define document metadata specifying the concepts which each document teaches, requires, and uses.

1. Introduction

Reusing available teaching resources is valuable as the resources are non-trivial and time consuming to create. The World Wide Web is potentially a source for a large number of teaching resources for a diverse range of topics. Reuse of those resources should allow for faster creation of a course and a richer coverage of areas of the course.

For example, when creating a course to teach the C++ Standard Template Library, there are many high-quality and varied teaching resources which can be reused. These range from the “Standard Template Library Programmer's Guide”1, to introductory tutorials such as “An introduction of STL for beginners”2, and entire books such as “Thinking In C++”3.

The potential value of these pre-existing resources is increased if the teaching system has some form of ontology for the domain and the resources have suitable markup in terms of that ontology. This can then be exploited by a teaching system which has a model of the student’s knowledge and their learning goals within this ontology.

For example, in the C++ Standard Template Library domain a student may wish to learn about the concepts ‘unique’ algorithm. If the teaching system has an ontology that indicates that the concepts ‘iterator’ and ‘remove’ are prerequisite knowledge for the concept ‘unique’ and the system believes the student understands the concept ‘iterator’ but not the concept ‘remove’, then it can create a teaching plan starting with the concept ‘remove’ and then moving to ‘unique’.

The creation of such an ontology is non-trivial as it involves enumerating the concepts and relationships between the concepts being taught. Many methodologies have been used for representing such domain knowledge, ranging from rule based expert systems to genetic graphs. In general, a domain expert must create the ontology in terms of the representation approach.

2. SITS Approach to Creating Ontologies

In SITS, the ontology is constructed from the document metadata. A course in SITS requires the definition of the vocabulary of concepts; each document in SITS has metadata specifying the concepts that the learner needs to know if they are to understand the document (prerequisites), the concepts the document teaches (learning outcomes), and the concepts it uses (where the use of these concepts does not require the learner understands them, as is the case for prerequisites). We refer to these three types as prerequisites, shows, and uses.

SITS uses this metadata to construct the inherent ontology for the course composed of the collection of documents put together by the creator of the learning environment. This ontology is dependent on the teaching strategy. For example, one teaching strategy might treat uses-metadata in the same way as prerequisite-metadata. This would result in a different ontology from that generated using a teaching strategy which did not treat uses-metadata in such a restrictive way. Teaching Strategies in SITS are separate modules and their implementation is not restricted, students select the teaching strategy they wish to use, and can change their selection at any time.

For a student wishing to learn about the concept ‘unique’ algorithm, SITS examines the metadata of the documents which show ‘unique’. Suppose we have the following pair of documents that teach the same concept:

1 http://www.sgi.com/tech/stl
2 http://www.mindcracker.com/mindcracker/c_cafe/stl/stlt1.asp
3 http://www.mindview.net/Books/TICPP/ThinkingInCPP2e.html
Suppose SITS believes the student understands the following concepts: ‘iterator’, ‘binarypredicate’, ‘vector’, ‘remove’, and ‘copy’. A teaching strategy, which treated uses concepts as non-essential could select either document to use to teach the ‘unique’ concept. The decision would be dependant on other factors (such as the extra concepts in shows for the second document, and the other metadata such as author and style). A teaching strategy which treated the concepts in uses as if they were prerequisites would have to first teach the student those concepts and in effect would be choosing between two ontologies:

\[(\text{iterator, remove, binarypredicate, vector, copy, sort}) \rightarrow \text{unique}\]

\[(\text{iterator, binarypredicate, container, sort, vector, generate, set}) \rightarrow \text{unique}\]

SITS represents student knowledge using an evidence based student model in which the concepts used in the document metadata are the items for which evidence of knowledge is collected.

3. SITS Evaluation

The main testing domain for SITS is the C++ Standard Template Library (STL). The course has approximately 200 concepts and 600 documents. We have created metadata for documents from four independent sources: an introductory, tutorial-style on-line textbook; parts of a higher level textbook; the standard C++ reference materials; and a collection of small, focussed examples which was developed for testing for STL implementations.

We have begun artificial experiments, with a range of different teaching strategies and we are experimenting in combining these with various student models. We compare the outcome of this SITS minimalist approach against expert definitions of good selections of teaching documents. We are also assessing the quality of the ontologies that are inferred from the metadata.

The ontologies are to be assessed by comparing the order of concept introduction they result in with the order used by a number of textbooks in the area, as well as comparing the ontologies directly with hand crafted ontologies of the area, and by having an expert in the area examine the generated ontology itself.

4. Conclusions

It is important to note that SITS’ metadata is the representative of the document in the system. In combination with the teaching strategy, it is used to infer the ontology for the domain.

In addition, SITS makes use of other metadata, such as the author of a resource and the style of the resource (eg. tutorial, reference, exercise, etc). This is imported from metadata sources, which are independent of SITS and are external to it. This forms an independent ontology of learning object type.

By contrast, the metadata, which relates the resource to the concepts being taught is part of SITS and its creation is the essential task for adding a new resource to a SITS teaching environment. It provides this mechanism that relates resources to concepts. In general, a SITS learning environment is created by defining the set of concepts before any documents are added to the environment. A major attraction of this approach is that the addition of the hundredth or thousandth resource is as easy as adding the first.

The one change that one expects over time is that teachers will become aware of the need for additional concepts after they have marked up some of the documents. SITS allows the creation of new concepts at any stage. This poses problems since such concepts could not be part of the metadata of documents already in the learning environment. Ideally, the teacher would revisit the coding of all existing documents to ensure that new concepts have been added to their metadata where this is relevant. However, at worst, the new concepts can simply be used for future document’s metadata.

SITS is being tested with the quite substantial STL domain. It remains to perform experiments to evaluate the quality and effectiveness of the ontologies generated by SITS. The SITS approach seems promising for the automatic and pragmatic construction of ontologies from modest collections of metadata that a teacher creates for each of the documents that are to be used or reused for a rich teaching environment.