Ontology Extraction for Educational Knowledge Bases *

Kyle Rawlins, Chris Eliot, Victor Lesser and **Beverly Woolf**
140 Governor's Drive, Department of Computer Science
University of Massachusetts, Amherst, MA 01003-4610, USA
{rawlins, eliot, lesser, bev}@cs.umass.edu

Introduction

A student who wishes to learn about some particular topic does not have many options. An often used tool is the search engine, which gives a tiny and difficult to control window into the vast amounts of information that is available on the internet. A student who wants to learn some concept should be able to interact with the available information in a coherent and personalized way. The classroom is the ideal of this goal, and our system would not replace, but augment it. It is within the reach of modern tutoring systems to use both knowledge of the student and of the subject's structure in order to present a subject in a manner that is more coherent and pedagogically sound than currently existing technology. One of the basic building blocks of such a system is the model of topic structure, and most importantly, how to obtain the information that fills the model.

Here we outline our research platform for the study of ontology lifecycle management, as well as several techniques that have so far had qualitative success. This research is taking place within the context of the Digital Libraries Initiative, under which thousands of instructional objects are organized, ranging from multimedia tutors (Woolf & Hall 1995), to lecture notes and papers. Our long term goal is to develop agent based tutoring systems which draw on this large knowledge base, and we have discussed our approach to this in other recent work (Eliot, Woolf, & Lesser 2001) (Woolf, Eliot, & Klein 2001) (Woolf *et al.* 2000).

Available online course materials range from short syllabi, to detailed breakdowns of the course with syllabi, lecture notes, and sometimes even textbooks online. Often, this information exists in disparate formats, and often contain large segments of free text or prose. In order for agents to draw effectively on this collection of information, we would like to find knowledge structure (ontology) that exists at a deeper level within the collection. This is a synthesis of data mining, extraction and fusion. It is embedded this second perspective that we are approaching the research, and this lends itself well to a multi-agent perspective.

Our knowledge structure is implemented in terms of an ontology represented in a database. We take this to be a structure consisting of *topics* or (interchangably) *concepts* and the *relationships* between these topics.

For our purposes, the internal structure of a topic differs from the common formulations in existing computational work on ontologies (Fellbaum 1998) and ontology extraction (Hearst 1992) (Caraballo 2001) (Goldman, Langer, & Roschenschein 1997). In these applications the goal is often much more lexically oriented, or more domain driven. In our system, the goal is domain driven but fairly abstract in that we want to represent ontological relationships that deal with learning. It would be more correct to refer to our notion of a concept as a *pedagogical concept*. This notion is compatible with but more specific than the general notion of concept in computational or philosophical ontology.

The data sources that we plan to use initially are course materials mostly found online, and consist of things like course syllabi, textbook chapters, lecture notes, etc. These are found in unstructured form (as simple web pages, usually, or text), and our goal is to discover structure from their text. We have done a certain amount of theoretical research as to how this might be done, as well as some implementation of basic ideas.

Basic assumptions One of our most important assumptions is that there is a conventional or normative pedagogical structure that is assumed about certain materials, rather than there being any absolutely true structure. Instead of telling educators what to do, we are attempting to describe what they do, and how they implicitly relate topics.

For example, we assume that when different instructors write a syllabus or textbook for a Data Structures class, there are patterns that are independent of the instructor. These patterns follow from an consensus as to how the topic should be taught. This consensus is not the 'best' way to teach something, but rather a convention. Embedded in such conventions is deeper structure of how knowledge is related.

*This material is based in part upon work supported by the National Science Foundation under Grant No. IIS-9812755. This work was supported in part by the Center for Intelligent Information Retrieval and in part by the National Science Foundation Cooperative Agreement number ATM-9732665 through a subcontract from the University Corporation for Atmospheric Research (UCAR). Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

Overview of the extraction process

On the level of a single agent performing maintenance on the ontology, there are three broad necessary behaviors. The first two are discovery of topics and discovery of relationships. These discovery tasks operate both on new and existing ontology. The third task is the process of analysing the existing knowledge structure in terms of newly discovered topics and relationships. The existing structure is taken as a hypothesis that is to be tested against newly discovered information. These tasks are independent, and each generates new information that is useful to the other two.

Evaluation An ontology for learning is subjective, no matter how it is constrcuted. It makes claims about how courses are taught and how information is learned, and we know of no gold standard that is any less subjective than an automatically generate ontology. Still, we are working on a small hand-generated ontology to attempt some evaluation of this sort.

Additionally, we are checking for internal consistency of the process by use of held-out data. That is, the extraction process is run many times from scratch with some randomly selected courses are held aside, and results of these successive runs are compared. This reflects on how sensitive our processes are to the input data, as well as how general the knowledge they are extracting is.

Topics

The topic is the basic unit of our ontology, and is better thought of as a *pedagogical topic*. At this stage, we are not able to give a theoretical definition of this notion, since that would be a difficult philosophical task. However, we can give intuitive and functional definitions which will suffice for practical purposes.

A topic is some unit of teaching. The granularity or level of abstractness can be quite varied, and can range from the topic of ”computer science”, down to the granularity of ”linked list” or ”pointer”. These both are parts (meronyms) of ”computer science” and of many intermediate topics. Topics are also either typically general categories, or instances of these categories meant for a specific course.

Our working definition of a topic is something that a person might attempt to teach or to learn. We will not pretend to offer a philosophically satisfactory criterion for knowing about a topic, but we feel comfortable judging in specific cases that a student does not know, or needs to know, or wants to know a topic. A teacher or tutor cannot be entirely confident without some assessment mechanism that any instruction has succesfully caused learning, but we can at least name specific topics that are the intended goal of a lecture or tutor.

A general formula for education relates prerequisites of concepts, study, and learning. If the prerequisites are satisfied, and study is truly completed, then learning of a topic is achieved. While terms in this formula are somewhat immeasurable, the structure can be taken as axiomatic in principle. Failure to learn can be blamed on failure of study, or unmet prerequisites. The relationships implied by this formula can be taken as constraints on the definition of a topic.

Derived topics Topics in our ontology can be taken from a number of sources. They could be directly entered by a human, extracted from a course syllabus (or other material), or derived from within the ontology itself. It is this last class that is the most important.

Topics taken directly from a syllabus or table of contents are directly inserted into the ontology, but alone don’t provide much information. They are almost uniformly what we consider to be non-abstract topics: specific instantiations of some unknown abstract concept class or category. The more interesting topics will be the ones that are derived from these instances, which are more abstract, general, and represent a teaching concept rather than some use of that concept.

For example, the topic ”linked lists” as found in a syllabus is taken as an instance of the general subject of linked lists. Instantiations of a category may have slightly different details and overlap, but are all referring to the same thing, their category.

We are using a derived topic that approximates a true category, a topic aggregate. When several courses on roughly the same subject material are added to the database, along with these come many topic instances that are closely related, having some kind of similarity relationship. For instance, many Data Structures courses will have a node for ”linked lists” in their syllabus, though the position, lecture contents, and wording of that node might vary. A topic aggregate is meant to represent the abstract version of these topics. Every topic in an aggregate is treated as being similar to every other topic in the aggregate. This can be thought of as clusters of similar topics, and the most direct techniques for finding such aggregates are in fact information retrieval based clustering algorithms.

Relationships

Just as important (or more) than the actual concepts in the ontology are the relationships between these concepts. There are many ”classic” ontology relationships that apply to our domain, as well as some that are more specific.

Relationships have certain general properties, many of which can be thought of in algebraic terms, since they are essentially relations on pairs of topics. A relationship could be transitivity, reflexivity, symmetricity, etc. For derived relationships, it is also useful to keep some notion of ’strength’ of the relationship, since most techniques we use at this point are probabilistic and do not produce results best represented with a relationship being on or off. Strength in this sense is difficult to define uniformly and the meaning of this value varies depending on the relationship and how it was derived.

At this point in time we have concentrated mostly on extracting just two relationships; one more general, and one that is domain-specific.

Equivalence and similarity The first relationship we approached started as an *equivalent-to* type of relationship (synonym). Two concepts would be synonyms if and only if they could be freely exchanged for each other in some (or all) contexts. This sort of relationship has some nice properties including being completely transitive and symmetric.

We found, however, that equivalence was not as useful as

we expected. At this level of abstraction, true synonyms are comparatively rare but things which seem ”similar” to each other are quite common, and we really want to also describe this second class. For example, far more common than finding two instances of a concept that is simply ”arrays” in a given course is finding concepts that are more like ”arrays of arrays” or ”Multidimensional arrays”.

In response to this, we started using the notion of a *similar-to* relationship. This for the most part works like the equivalence relationship (with similar properties of opacity), except that it is much less theoretically clean. It is not accurate to say that *similar-to* is transitive or even necessarily symmetric. While two particular topics might be similar, a large ontology might have long chains of similarity. If this were transitive very unlikely things would be considered ”similar”.

To combat this issue we have introduced a notion of partial transitivity, where the transitivity falls off a certain amount across a relationship as a factor of the strength of the relationship. It is also intuitively desirable to give some weight to a similarity relationship, so to say that some topics are more or less similar. This can be thought of as something like percentage of similarity, but is really an approximation of some (undiscoverable) ideal probability.

Extraction of similar-to Currently we are using a standard hierarchical clustering algorithm that operates on just the text of a concept. This works reasonably well for small data sets, but may have some problems scaling. The clustering occurs after some processing, including stemming and normalization.

There are some problems with this despite its simplicity. With a small data set it is fairly easy to terminate the clustering algorithm and pick the correct number of clusters, since they simply stop merging at a certain point, so the use of our clustering algorithm requires more study. Also, the clustering is occurring based on word content rather than any notion of semantic content. For our purposes it seems the second is much more ideal, but is unfortunately much harder. One line of approach would be to use WordNet or something like it as a base for semantic analysis.

Concept prerequisites When learning a particular topic, there are likely to be concepts that would be helpful or even necessary to understand. These can be thought of as concept prerequisites. This is directly valuable information because it gives clues about how one would generate a ’learning sequence’ between two arbitrary topics. A related relationship is temporal orderings between two topics in a learning sequence. Such an ordering does not always signal a real concept prerequisite, but can be used as an indication of one. Evidence of what these orderings are is easily found, since most course materials have a somewhat linear organization.

There is clearly some notion of transitivity that is applicable here, though the specifics are not so clear. If A must be learned before B, and B must be learned before C, then it is clear that A must be learned before C. However, it is not necessarily the case that A should be considered a concept prerequisite for C; the issue is one of opacity.

Extraction of temporal orderings There are two techniques that we have looked at for inferring this information. The first consists of looking and seeing whether, on average, instances of two abstract topics are before or after each other in courses where the instances are both found. If they have some ordering which is consistent across many courses, that ordering is added to the ontology as a temporal ordering requirement. If they are not consistent (in some one is after, in some it is before), the information is valuable (they have some kind of independence) but is not currently dealt with.

Our second approach treats the course syllabi as emissions of some markov model with unknown transition probabilities, and attempts to learn these probabilities. The important orderings in the resulting markov model are the transitions with high chance that are learned during training.

Conclusion and future directions

We have given a broad overview of our ideas and some of the technology that we have implemented, but many of the techniques are still only in their infancy. In particular, the later stages of the management cycle, where the results of the data fusion are evaluated with respect to the existing ontology are still quite undeveloped. The next major step is to complete and analyze experimental results for our techniques we are developing, and this is our current task. Already the results are qualitatively promising, and we are close to quantitative evaluations of the ontology management.

References

Caraballo, S. A. 2001. *Automatic construction of a hypernym-labeled noun hierarchy from text*. Ph.d. diss, Brown University.

Eliot, C.; Woolf, B.; and Lesser, V. 2001. Knowledge extraction for educational planning. In *Workshop on Multi-Agent Architectures Supporting Distributed Learning, at the 10th International Conference on Artificial Intelligence in Education*.

Fellbaum, C., ed. 1998. *WordNet: an electronic lexical database*. Bradford Books.

Goldman, C.; Langer, A.; and Roschenschein, J. S. 1997. Musag: An agent that learns what you mean. *Journal of Applied AI, a special issue on Practical Applications of Intelligent Agnets and Multiagent Technology* 11(5):413–435.

Hearst, M. A. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the fourteenth international conference on computational linguistics*.

Woolf, B., and Hall, W. 1995. Multimedia pedagogues: Interactive multimedia systems for teaching and learning. *IEEE Computer* 28(5):74–80.

Woolf, B.; Lesser, V.; Eliot, C.; Eyeler-Walker, Z.; and Klein, M. 2000. A digital marketplace for education. In *International Conference on Advances in Infrastructure for Electronic Business and Education*.

Woolf, B.; Eliot, C.; and Klein, M. 2001. *A Digital MarketPlace for Education*. Norwell, MA: Kluwer Academic.