Ontology Extraction for Educational Knowledge Bases*

Peter Cassin, Chris Eliot, Victor Lesser, Kyle Rawlins, and Beverly Woolf

140 Governor's Drive, Department of Computer Science
University of Massachusetts, Amherst, MA 01003-4610, USA
{pcassin, eliot, lesser, rawlins, bev}@cs.umass.edu

1 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 life-cycle 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 [1], 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 [2, 3, 4].

There are two main components to this paper. First, we describe our architecture for extracting structured information from raw web pages. Second, we describe our techniques for extracting a more complete ontology of pedagogical information from the structured information.

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

* 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.

L. van Elst, V. Dignum, and A. Abecker (Eds.): AMKM 2003, LNAI 2926, pp. 297–309, 2003.