Discovering and Building Semantic Models of Web Sources*

Craig A. Knoblock

University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292 USA
knoblock@isi.edu

Abstract. To achieve widespread use of the Semantic Web depends on having a critical mass of Web data available with semantic annotations. Since there are a huge number of sources available today without any such annotations, the challenge is how to find and build semantic models for these sources. In this talk I will describe an integrated end-to-end approach that automatically discovers information-producing web sources, invokes and extracts the data from these sources, builds semantic models of the sources, and validates the results by comparing the data produced by the source with the model of the source. These techniques are implemented in a system called DEIMOS, which integrates a diverse set of technologies to completely automate this task. DEIMOS starts with a "seed" source and finds other similar sources online using data from a social networking web site. Next the system learns how to invoke these sources through experimentation and then extracts data from these sources with automatic wrapping techniques. Finally, DEIMOS learns a semantic model of a source, which identifies the semantic types of the data produced by a source as well as the function that maps the inputs to the outputs. I will describe the challenges in integrating the component technologies into a unified approach to discovering, extracting and modeling new online sources. I will also present an evaluation of the integrated system on three different domains to demonstrate that it can automatically discover and model new Web sources.

* This talk is based on joint work with Jose Luis Ambite, Mark Carman, Cenk Gazen, Kristina Lerman, Steven Minton, Anon Plangprasopchok, and Tom Russ. This research is based upon work supported in part by the National Science Foundation under award number IIS-0535182, in part by the Air Force Office of Scientific Research under grant number FA9550-07-1-0416, and in part by the Defense Advanced Research Projects Agency (DARPA) under Contract No. FA8750-07-D-0185/0004.

L. Aroyo et al. (Eds.): ESWC 2009, LNCS 5554, p. 2, 2009.