Discovering and Building Semantic Models of Web Sources [1]

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.

About Craig Knoblock

Dr. Craig Knoblock is a Senior Project Leader at the Information Sciences Institute and a Research Professor in Computer Science at the University of Southern California (USC). He is also the Chief Scientist for both Fetch Technologies and Geosemble Technologies, which are spinoff companies from USC.
He received both his M.S. and Ph.D. in Computer Science from Carnegie Mellon and his B.S. from Syracuse University.
His current research interests include information integration, information extraction, machine learning, users interfaces, constraint reasoning, geospatial data fusion, and bioinformatics. He

has published one book and over 150 articles, book chapters, and conference papers on his research. He has served on the Senior Program Committees of the National Artificial Intelligence Conference (1997, 1998, 2000, 2004, 2006, 2007), the International Joint Conference on AI (2007), the International Semantic Web Conference (2004, 2008), and the International Conference on Intelligent User Interfaces (2009).
He was program co-chair for the 2008 AAAI track on AI and the Web and he is conference chair for the 2011 International Joint Conference on AI (IJCAI). He is on the editorial board of AAAI Press, Computational Intelligence, and the Journal on Foundations and Trends in Web Science. He is a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI), a Distinguished Scientist of the Association of Computing Machinery (ACM), a Trustee of the International Joint Conference on Artificial Intelligence (2007-2017), and past President of the International Conference on Automated Planning and Scheduling (2006-2008).

back

---

[1] 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.