Information Integration on the Web

Subbarao Kambhampati, *Arizona State University*

Craig A. Knoblock, *University of Southern California*

This special issue contains articles based on a sampling of the presentations from the two-day Workshop on Information Integration on the Web, held in August 2003 at the International Joint Conference on Artificial Intelligence, in Acapulco, Mexico. IIWeb '03 brought together researchers working in a variety of areas related to the

1094-7167/03/$17.00 © 2003 IEEE
Published by the IEEE Computer Society

larger problem of integrating information on the Web. These areas include machine learning, data mining, automated planning, constraint reasoning, databases, data integration, information extraction, the Semantic Web, and Web services.

Effective integration of heterogeneous databases and information sources has been cited as the most pressing challenge in spheres as diverse as corporate data management, homeland security, counter-terrorism, and the human genome project. This issue's articles are representative of the challenges that information integration research is tackling. (For a quick tutorial on information integration, see the slides from our AAAI 2002 tutorial, at http://rakaposhi.eas.asu.edu/i3-tut.html.)

Data extraction

Integrating information from Web sources often starts by extracting the data from the Web pages exported by the data sources. Although XML is supposed to reduce the need for this extraction, relatively few sources are currently available in XML, and legacy HTML sources will be around for years to come. In "Reconfigurable Web Wrapper Agents" in this issue, Chia-Hui Chang and her colleagues present a tool for rapidly generating agents for extracting data from Web sites. The tool supports programming by example. A human indicates which parts of the HTML text correspond to which data fields, and then the tool automatically generates a program for extracting the data.

Deriving semantics

Because of the semantic heterogeneity among sources, merely extracting the data from Web pages is often insufficient to support integration. The problem is that information might be organized in different ways with different vocabularies. So, an integration system needs to either learn or have access to semantic descriptions of the sources. This can be done by either bundling semantic information with Web pages or learning ontologies from the sources. Two articles in this issue attempt to derive semantics associated with individual pages. In "OntoMiner: Bootstrapping and Populating Ontologies from Domain-Specific Web Sites," Hasan Davulcu and his colleagues introduce an approach for automatically extracting ontologies from Web pages. In "Annotation for the Deep Web," Siegfried Handschuh, Steffen Staab, and Raphael Volz describe an approach for automatically adding semantic annotations for Web pages that are generated by a back-end database server.

Relating data from different sources

Once a system can extract information from the various sources and has a semantic description of these sources, the next challenge is to relate the data in the sources. Many sources use different ways to describe the same entities or objects. To integrate data across sources, an integration system must be able to accurately determine when data in two different sources refer to the same entities. Two articles in this issue address this *name matching* or *object matching* problem. "Adaptive Name Matching in Information Integration," by Mikhail Bilenko and his colleagues, surveys methods for matching object names across different data sources. In "Profile-Based Object Matching for Information Integration," AnHai Doan and his colleagues outline a method for matching objects from different data sources using differing vocabularies. Their method uses the mappings between the relation and attribute names among the schemas of the individual data sources to help determine the object mappings.

Query processing

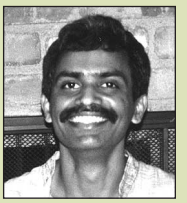
After the mapping between data sources is completed, the next issue is how to effectively reformulate a user query into queries on individual data sources. Effective reformulation strategies must be sensitive to the data sources' constraints to access the smallest number of most relevant sources when answering the query. "Using Constraints to Describe Source Contents in Data Integration Systems," by Chen Li, describes approaches for modeling and using different types of source constraints.

Architectures

Although most of the articles assume a centralized "mediator" architecture for data integration, an important alternative is a peer-to-peer architecture. In this architecture, a set of data sources can act as both sources and mediators. Peer-to-peer architectures, although flexible, introduce a variety of issues in schema mapping and query reformulation. "Querying Distributed Data through Distributed Ontologies: A Simple but Scalable Approach," by François Goasdoué and Marie-Christine Rousset, addresses some of these issues.

This special issue covers only a few of the many active research projects dealing with information integration on the Web. Owing to space limitations, we can't cover the other topics presented at IIWeb '03. Such topics include matching schemas or ontologies across sources, gathering statistics about the sources, efficiently executing information-gathering plans, and information integration in bioinformatics. For the complete set of workshop papers, access www.isi.edu/info-agents/workshops/ijcai03/proceedings.htm. ■

The Authors



Subbarao Kambhampati is a professor of computer science and engineering at Arizona State University, where he directs the Yochan research group. His research interests span topics in AI and databases, and include automated planning and information integration. He received his PhD from the University of Maryland, College Park, and was the recipient of a 1994 National Science Foundation young investigator award. Contact him at the Dept. of Computer Science and Eng., Arizona State Univ., Tempe, AZ 85287-5406; rao@asu.edu; http://rakaposhi.eas.asu.edu/rao.html.

Craig A. Knoblock is a senior project leader at the University of Southern California's Information Sciences Institute and a research associate professor in USC's Computer Science Department. His research interests involve developing and applying planning, machine learning, and knowledge representation techniques to the problem of information gathering and integration. He received his PhD in computer science from Carnegie Mellon University. Contact him at the Information Sciences Inst., Univ. of Southern California, 4676 Admiralty Way, Marina del Rey, CA 90292-6696; knoblock@isi.edu; www.isi.edu/~knoblock.