Information Integration on the Web: A view from AI and Databases (Report on IIWeb-03)*

Subbarao Kambhampati
Dept. of Computer Science and Engineering
Arizona State University
Tempe, AZ 85287-5406

Craig A. Knoblock
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292.

Introduction

This document is a report on the workshop on Information Integration on the Web (IIWeb-03), held in Acapulco, Mexico, on August 9-10, as part of the 2003 International Joint Conference on Artificial Intelligence. The full proceedings of the workshop are available online [1]. A small sample of the papers presented at the workshop were also included in a special issue of IEEE Intelligent Systems [2].

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. An important impediment to scaling up integration frameworks to large-scale applications has been the fact that the autonomous and decentralized nature of the data sources constrains the mediators to operate with very little information about the structure, scope, profile, quality and inter-relations of the information sources they are trying to integrate. As stated, the problem of information integration[1] crosses the boundaries of AI and Databases, and includes research in the areas of machine learning, data mining, automated planning, constraint reasoning, databases, view integration, information extraction, semantic web, web services, and other related areas.

Not surprisingly, the problem of information integration has drawn significant interest from both AI and Databases. Although there are a variety of forums where research on information integration is presented, most of these forums are naturally seen as "belonging" to either the Artificial Intelligence (AAAI, IJCAI, ICML, etc.) or Database (SIGMOD, VLDB, ICDE, CIKM, etc.) communities. The primary purpose of this workshop was thus to bring together researchers from AI and Databases who are working in a variety of problems related to integrating information on the Web.[2]

Structure and Organization

At the outset, we were fortunate to be able to convince a diverse group of researchers from both the AI and DB communities to help us in organizing this workshop. The workshop call for papers had a very good response. We received 40 submissions spanning a diverse set of issues relevant to information integration. After reviewing, 18 papers were scheduled for oral presentation and another 15 were scheduled for poster presentation.

With close to 50 attendees, ours was one of the larger (and livelier) workshops at IJCAI. To encourage discussion, the oral presentations were structured into topic-oriented panels. In each panel, the authors had 20 minutes to present their papers, followed by a 30 minute panel discussion on the topic. In addition to the contributed papers, there were two invited panels (see below). Finally, thanks to the sponsorship of RIACS, the attendees had on-site catered lunches, which facilitated many lively off-line discussions.

Technical Sessions

There were three topic panels on the first day and three panels on the second. The panels were organized in the rough sequence of sub-problems that arise in information integration. Integrating information from Web sources often starts by extracting the data from the Web pages exported by the data sources. Accordingly, the first session of the workshop ("*Wrapping and Extracting*") had three papers on extracting information from text-based web 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

*IJCAI 2003 Workshop on Information Integration on the Web

[1] For a tutorial introduction to issues in Information Integration, the readers are referred to [3].

[2] Earlier meetings that had similar goals include: the 1995 AAAI Spring Symposium on Information gathering; as well as the 1998 and 1999 AAAI and IJCAI workshops on Intelligent Information Integration.

two different sources refer to the same entities. The second session of the workshop ("*Name Matching*") had two papers on strategies for matching names and objects across data sources. The third session ("*Schema Matching*") had three papers on matching and integrating schemas across data sources.

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. A related issue is that of gathering statistics about the organization and distribution of data across sources. Given that Web sources are autonomous, probing and learning techniques have to be used to gather statistics. The first session of the second day ("*Meta-data and Statistics*") had four papers related to these issues.

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, and how to execute the resulting queries. The session titled *Query Processing and Execution* was devoted to these issues. In addition to query processing and execution, this session also had a paper on integration challenges involved in peer-to-peer databases, and in managing the data inconsistency across sources.

Although much of the early work on information integration was driven by business applications, more recently scientific data integration tasks have become the driving applications for this work. Prominent among these is the task of integrating the public domain bioinformatic data sources. The workshop had a session devoted to this task, where two papers were presented. The session produced lively discussion about the steep setup costs involved in familiarization with bioinformatic sources.

Panels

In addition to the topic panels, the workshop had two panels, one at the end of each day. The panel on the first day was titled "*The Economics of Information Integration: The Practical View of Information Integration on the Web*." The panelists included representatives from current and former companies involved in Information Integration with Steven Minton from Fetch Technologies, David Pennock from Overture, Amir Ashkenazi from DealTime, and Andrew McCallum previously from WhizBang. The panel on the second evening was titled "*Future Funding For Information Integration*" and the panelists included Michael Pazzani from NSF and Barney Pell from NASA. Michael Pazzani provided an overview of the NSF/ITR program for the next year, which will have Information Integration as one of the themes. Dr. Pazzani discussed the compelling importance of *ad hoc* information integration. Barney Pell gave an overview of the Information Integration components of the NASA Discovery Systems Program, which is expected to start in 2004/2005.

Prognosis

The general consensus of the participants at the workshop was that the meeting was successful in improving the connections between AI and DB researchers working on Information Integration. The attendees felt that it would be worth continuing the workshop in subsequent years–perhaps alternating its location between premiere AI and Database conferences.

Acknowledgements

We would like to thank the members of our organizing committee (Lise Getoor, Alon Halevy and Sheila McIlraith); and the members of our program committee (William Cohen, Hasan Davulcu, Anhai Doan, Juliana Freire, C. Lee Giles, Joseph M. Hellerstein, Nicholas Kushmerick, Andrew McCallum, Giansalvatore Mecca, Renee Miller, Ami Motro, Jeffrey Naughton, Louiqa Raschid, Marie-Christine Rousset, and Sheila Tejada) for their many invaluable inputs and thoughtful reviews. We would also like to thank Serdar Uckun of Research Institute for Advanced Computer Science (RIACS) for providing financial support for the workshop.

References

[1] Online proceedings of the IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03). URL: *www.isi.edu/info-agents/workshops/ijcai03*

[2] Information Integration on the Web. Special issue of IEEE Intelligent Systems, 18(5), September/October, 2003. URL: *computer.org/intelligent*.

[3] Information Integration on the Web. Craig Knoblock and Subbarao Kambhampati. Tutorial presented at AAAI 2002. URL: *rakaposhi.eas.asu.edu/i3-tut.html*.