

Workshop on Intelligent Information Integration (III'99)

Dieter Fensel, Craig Knoblock, Nicholas Kushmerick,
and Marie-Christine Rousset

The growth of the internet and other online information repositories has greatly simplified the access to numerous sources of information. However, it has become cliché to observe that this growth has vastly complicated tasks that involve finding, extracting, merging, or synthesizing such information. With today's tools, a tourist planning a trip to Paris cannot simply fetch from the web a map showing all Italian restaurants close to the Eiffel Tower, although this information might well be available on the web if maps of Paris and lists of restaurants and tourist sites are combined. The problem is not one of information distribution but one of information integration. Building tools to simplify access to the wealth of available information constitutes a significant challenge to computer science. AI is well suited to this challenge, with its emphasis on heuristic solutions, knowledge representation and knowledge-intensive techniques, explicit management of uncertainty, learning and adaptivity, planning, and so on. Over the past several years, the information-integration community in AI has been exploring these issues. The result has been substantial progress on problems such as representing information sources; planning information-gathering actions; handling semantic heterogeneity; optimizing queries; caching; exploiting ontological knowledge; developing tools for web site construction; discovering resources; automatically learning to access sources; and scaling information-gathering systems to the real world.

Roots of the Field and Major Approaches

Work on information integration began as soon as heterogeneous and distributed information sources had to be combined to process tasks successfully. A landmark paper that still guides most of the current research is Wiederhold (1992) (a recent outlook paper by the same author has appeared as Wiederhold [1997]). He identified the need for a middleware architecture that buffers the heterogeneous and distributed information sources at the one side and the infor-

The Workshop on Intelligent Information Integration (III), organized in conjunction with the Sixteenth International Joint Conference on Artificial Intelligence, was held on 31 July 1999 in Stockholm, Sweden. Approximately 40 people participated, and nearly 20 papers were presented. This packed workshop schedule resulted from a large number of submissions that made it difficult to reserve discussion time without rejecting an unproportionately large number of papers. Participants included scientists and practitioners from industry and academia. Topics included query planning, applications of III, mediator architectures, and the use of ontologies for III.

mation client at the other side. In fact, this middle layer again has two distinct interfaces reflecting its mediating purpose. *Wrappers* abstract from the syntactic idiosyncracies of the information sources, translating sources' native data formats into a representation that can be used for integration

(Muslea, Minton, and Knoblock 1999; Kushmerick 1997). The *mediator* interacts with the information client and translates his/her queries into subqueries passed to the various wrapped information sources. Numerous projects have been implemented in this architecture; examples include CARNOT (Collet, Huhns, and Shen 1991), HARVEST (Bowmann et al. 1994), HERMES,¹ INFOMASTER (Genesereth, Keller, and Duschka 1997), INFORMATION MANIFOLD (Levy, Rajaraman, and Ordille 1996), MOMIS (Bergamaschi et al. 1998), OBSERVER (Mena et al. 1996), SIMS (Arens, Knoblock, and Shen 1996), and TSIMMIS (Papakonstantinou, Garcia Molina, and Widom 1995).

Originally, the field was mainly concerned with integrating various databases. For example, the semantic values of Sciore, Siegel, and Rosenthal (1994) describe an approach specialized for the integration of relational databases. The increasing popularity of the web with its huge amount of semistructured information has significantly changed the focus. Here, information is weakly structured, heterogeneous, and distributed by nature. However, the medium covers most information needs if one is able to find the information one is looking for. Approaches such as SHOE (Luke et al. 1997) and ON2BROKER (Fensel et al. 1999) are specialized for the integration of web sources with the use of ontologies. Originally, these approaches were hampered by poor web standards such as HTML, but new standards such as XML and RDF make semantic access to information much easier. Not surprisingly, many startup companies are arising in this area, and intelligent information integration (III) is ready to take over this marketplace.

Aim of the Workshop

In the past years, a number of workshops have been organized on III. There was a workshop during the German AI conference in 1997, a workshop at ECAI-98, and a workshop during the Fifteenth American Association for Artificial Intelligence (AAAI-98). This workshop, collocated with the Sixteenth International Joint Conference on Artificial Intelligence,

...the cutting edge in AI research...

Machine Learning

Executive Editor:

Robert C. Holte

University of Ottawa, Canada

Machine Learning is an international forum for research on computational approaches to learning. The journal publishes articles reporting substantive research results on a wide range of learning methods applied to a variety of task domains, including but not limited to:

Methods: Inductive learning methods; Explanation-based learning; Genetic algorithms; Analogy and case-based methods; Connectionist techniques; Automated knowledge acquisition; Learning from instruction.

Task Domains: Classification and recognition; Problem solving and planning; Reasoning and inference; Natural language processing; Design and diagnosis; Vision and speech perception; Robotics and motor control.

The ideal paper will make a theoretical contribution supported by a computer implementation. In addition to carefully describing the learning component, it should also discuss knowledge representation and performance assumptions. The article should carefully evaluate the approach through empirical studies, theoretical analysis, or comparison to psychological phenomena, and should discuss its relation to other work in machine learning. Variations from this prototype, such as critical reviews of existing work, will be considered provided they make a clear contribution to the field.



Subscription Information

2000, Volumes 38-41 (12 issues)

ISSN 0885-6125

Subscription Rate:

NLG 2013.00/USD 959.00, incl.p/h.

Private Rate:

NLG 960.00/USD 480.00, incl.p/h.

Special rate for AAAI/CSCSI: NLG 350.00/USD 175.00,

**P.O. Box 322, 3300 AH Dordrecht, The Netherlands
P.O. Box 358, Accord Station, Hingham, MA 02018-0358, U.S.A.**

<http://www.wkap.nl>

**Kluwer
academic
publishers**



shop was attended by about 40 participants.

Construction of Query Plans for Information Gathering

The first session of the workshop was on the construction of query plans for information gathering. Alon Levy, of the University of Washington, gave an invited talk entitled "Self-Organizing Data Sharing Communities," presenting his view on the future of III, where toasters, alarm clocks, video equipment, and other devices in a household negotiate their communication and share and exchange information about their human users. The papers by AnHai Doan (of the University of Washington) and Levy, "Efficiently Ordering Query Plans for Data Integration," Marc Friedman, Levy, and Todd Millstein (all of the University of Washington), "Navigational Plans for Data Integration," and Subbarao Kambhampati and Senthil Gnaprakasam (both of the University of Arizona), "Optimizing Source-Call Ordering in Information Gathering Plans," dealt with various aspects of how to optimize the computationally hard problem of answering user queries using decomposed subqueries on information sources.

Case Studies of Information-Integration Systems

The second session was on case studies of information-integration systems. The paper by Greg Barish et al. (all of the University of Southern California), "THEATERLOC: A Case Study in Building an Information-Integration System," described the design and implementation of an information-integration application that allows users to retrieve information about theaters and restaurants for cities in the Los Angeles area, including an interactive map depicting their relative locations and video trailers of the movies playing at the various area theaters. The paper by S. Bergamashi et al. (all of the University of Modena and Reggio Emilia), "Intelligent Techniques for the Extraction and Integration of Heterogeneous Information," introduced the language ODL for describing information sources by combining description logics and object orientation.

should bridge the gap between the European and American III communities. Some of the topics we proposed were the modeling of the contents of information systems, approaches for efficient and flexible query planning, the learning or constructing of wrappers for extracting data, the resolving of inconsistencies in identifying

objects, integration approaches using ontologies and contexts, languages for information integration, advanced integration architectures, information sources in the internet and the intranet, intelligent information retrieval, and federated databases and multidatabases. The call attracted about 30 submissions, and the work-

Finally, the paper by Mathias Bauer and Dietmar Dengler (both of DFKI), "TRIAS: Trainable Information Assistants for Cooperative Problem Solving," presented an approach to cooperative problem solving in which an information agent and its user try to support each other in the achievement of a particular goal.

Use of Ontologies

Ontologies have been developed by the knowledge engineering community to facilitate knowledge sharing and reuse, cf. Meersman (1999) and Gruber (1993). Naturally, they find large application potential for information sharing and integration. The paper by Yannis Labrou and Tim Finin (both of the University of Maryland), "Experiments on Using Yahoo! Categories to Describe Documents," reported on experiments with the Yahoo! ontology and the use for automatic document classification. The paper by Frank van Harmelen and Dieter Fensel (both of Vrije Universiteit Amsterdam), "Practical Knowledge Representation for the Web," discussed various means to represent meta-information of web documents (HTML-META, HTML-SPAN, CSS, XML, RDF). Finally, the paper by Stephen Crane-field and Martin Purvis (both of University of Otago), "UML as an Ontology Modeling Language," investigated the use of UML for representing ontologies, providing a link to tools and user habits from the software-engineering community.

Short Papers

A number of short papers covered various aspects of III. The paper by Mário Flecha (of the Universidade de Federal de Minas Gerais) and José Braga (of the Universidade de Federal de Viçosa), "Mediators to Access Public Data Resources," reported experience for database access in large and legacy real-world environments. The paper by Martin Speck and Christoph Uthmann (both of Westfälische Wilhelms-Universität Münster), "Web-Integrated Information Integration," is on information integration via web-based publishing work flows. The paper by Ichiro Ide et al. (all of the University of Tokyo), "Relating Graphical Features with Concept Classes for Automatic

News Video Indexing," proposed a new method for automatic indexing of video data. The paper by Thierry Poibeau (of Thompson-CSF), "Mixing Technologies for Intelligent Information Extraction," presented limitations of current information-extraction systems and showed how some of them can be bypassed by combining these systems. The paper by Elhadi Shakshuki (of the University of Waterloo), Hamada Ghenniwa (of Mackenzie Financial Corporation), and Mohamed Kamel (of University of Waterloo), "Information-Gathering System: Internet Navigation," presented a multiagent system that assists different users in locating, retrieving, and integrating information on the web. The paper by Jacques Calmet and Peter Kullman (both of University of Karlsruhe), "Meta Web Search with KOMET," presented a logic-based mediator system and its application to the web. The paper by Fensel et al., "ON2BROKER: Semantic-Based Access to Information Sources at the WWW," presented recent developments in the ON2BROKER projects, reflecting lessons learned from earlier versions of the system. Finally, Pepijn Visser and Valentina Tamma (both of the University of Liverpool) presented "An Experience with Ontology Clustering for Information Integration," discussing a structure of multiple shared ontologies to integrate heterogeneous information sources.

Conclusions

This workshop provided a platform to present, discuss, and evaluate current research and interesting applications of III. The field looks rather mature concerning its theoretical foundations as well as in the number of applications that exist. Many research projects are ongoing worldwide, and a need for further discussions and exchange of results exists. Some participants felt unhappy with the relatively small number of researchers from the database community. The participants discussed the possibility of organizing future workshops in connection with events of the database community in an effort to bring the two communities together.

The workshop proceedings are available at SunSITE. Informatik. RWTH-Aachen.DE/Publications/CEUR-WS/Vol-23/. Further information about the workshop can be found at www.aifb.uni-karlsruhe.de/WBS/dfe/iii99.html. General information on the III community and its mailing list can be found at www.tzi.de/grp/i3/.

Note

1. www.cs.umd.edu/projects/hermes/overview/paper/index.html.

References

- Arens, Y.; Knoblock, C. A.; and Shen, W.-M. 1996. Query Reformulation for Dynamic Information Integration. *Journal of Intelligent Information Systems* (Special Issue on Intelligent Information Integration) 6(2-3): 99-130.
- Bergamaschi, S.; Castano, S.; De Capitani di Vimercati, S.; Montanari, S.; and Vincini, M. 1998. An Intelligent Approach to Information Integration. In *Formal Ontologies in Information Systems*, ed. N. Guarino, 287-301. Frontiers in Artificial Intelligence and Applications. Amsterdam, The Netherlands: IOS.
- Bowman, C. M.; Danzig, P. B.; Manber, U.; and Schwartz, M. F. 1994. Scalable Internet Resource Discovery: Research Problems and Approaches. *Communications of the ACM* 37(8): 98-107.
- Collet, C.; Huhns, M. N.; and Shen, W.-M. 1991. Resource Integration Using a Large Knowledge Base in CARNOT. *IEEE Computer* 24(12): 55-62.
- Fensel, D.; Angele, J.; Decker, S.; Erdmann, M.; Schnurr, H.-P.; Staab, S.; Studer, R.; and Witt, A. 1999. ON2BROKER: Semantic-Based Access to Information Sources at the WWW. Paper presented at the World Conference on the WWW and Internet (WebNet 99), 25-30 October, Honolulu, Hawaii.
- Genesereth, M. R.; Keller, A. M.; and Duschka, O. 1997. INFOMASTER: An Information-Integration System. In Proceedings of the 1997 ACM SIGMOD Conference, 13-15 May, Tucson, Arizona.
- Gruber, T. R. 1993. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition* 5(2): 199-220.
- Kushmerick, N. 1997. Wrapper Induction for Information Extraction. Ph.D. dissertation, Department of Computer Science and Engineering, University of Washington.
- Levy, A. Y.; Rajaraman, A.; and Ordille, J. J. 1996. Query-Answering Algorithms for Information Agents. In Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96), 40-47. Menlo

Park, Calif.: American Association for Artificial Intelligence.

Luke, S.; Spector, L.; Rager, D.; and Hendler, J. 1997. Ontology-Based Web Agents. In Proceedings of the First International Conference on Autonomous Agents (Autonomous Agents97), 59–66, ed. W. L. Johnson. New York: Association for Computing Machinery.

Meersman, R. A. 1999. The Use of Lexicons and Other Computer-Linguistic Tools in Semantics, Design, and Cooperation of Database Systems. In *CODAS Conference Proceedings*, ed. Y. Zhang, 1–14. Lecture Notes in Artificial Intelligence (LNAI). Berlin: Springer Verlag.

Mena, E.; Kashyap, V.; Sheth, A.; and Illarmandi 1996. OBSERVER: An Approach for Query Processing in Global Information Systems Based on Interoperation across Pre-existing Ontologies. In Proceedings of the First IFCIS International Conference on Cooperative Information Systems (Coop-IS'96), 21–23 June, Brussels, Belgium.

Muslea, I.; Minton, S.; and Knoblock, C. A. 1999. A Hierarchical Approach to Wrapper Induction. In Proceedings of the Third International Conference on Autonomous Agents (Agents-99), 1–5 May, Seattle, Washington.

Papakonstantinou, Y.; Garcia Molina, H.; and Widom, J. 1995. Object Exchange across Heterogeneous Information Sources. In Proceedings of the IEEE International Conference on Data Engineering (ICDE), 6–10 March, Taipei, Taiwan.

Sciore, E.; Siegel, M.; and Rosenthal, A. 1994. Using Semantic Values to Facilitate Interoperability among Heterogeneous Information Systems. *ACM Transactions on Database Systems* 19(2): 254–290.

Wiederhold, G. 1997. Value-Added Mediation in Large-Scale Information Systems. In *Database Application Semantics*, ed. R. Meersman and L. Mark, 34–56. New York: Chapman and Hall.

Wiederhold, G. 1992. Mediators in the Architecture of Future Information Systems. *IEEE Computer* 25(3): 38–49.

Dieter Fensel is an associate professor at the Free University of Amsterdam in the area of electronic commerce. After studying mathematics, sociology, and computer science in Berlin, he joined the Institute AIFB at the University of Karlsruhe in 1989. His major subject was knowledge engineering, and his Ph.D. thesis in 1993 was about a formal specification language for knowledge-based systems. From 1994 until 1996, he visited the group of Bob Wielinga at the SWI Department in Amsterdam. During this time, his main interest was problem-

Computation & Intelligence

Collected Readings

EDITED BY GEORGE F. LUGER

This comprehensive collection of 29 readings covers artificial intelligence from its historical roots to current research directions and practice. With its helpful critique of the selections, extensive bibliography, and clear presentation of the material, *Computation and Intelligence* will be a useful adjunct to any course in AI as well as a handy reference for professionals in the field.

ISBN 0-262-62101-0 700 pp., index.

The AAAI Press / Distributed by The MIT Press

Massachusetts Institute of Technology, Cambridge, MA 02142

To order, call toll free:

800-356-0343 or 617-625-8569. MasterCard and VISA accepted.

solving methods of knowledge-based systems. In 1996, he came back as a senior researcher at the Institute AIFB, finalizing his habilitation in 1998. Currently, his focus is on the use of ontologies to mediate access to heterogeneous knowledge sources and apply them in electronic commerce. His e-mail address is dieter@cs.vu.nl.

Craig Knoblock is a project leader at the Information Sciences Institute and a research associate professor in the Computer Science Department at the University of Southern California (USC). He is also on the faculty of the Integrated Media Systems Center, which is a National Science Foundation Engineering Research Center at USC. He received his Ph.D. in computer science from Carnegie Mellon University in 1991 and joined USC that year. His current research interests include information agents, information integration, automated planning, and machine learning. He is one of the xPRIMARY architects of the SIMS information mediator, a system for integrating heterogeneous data sources. He jointly leads the ARIADNE and THESEUS projects, which are addressing the problems of building agents for integrating and managing web-based information sources. His e-mail address is knoblock@isi.edu.

Nicholas Kushmerick is a college lecturer in the Department of Computer Science at

the University College Dublin. He received his B.S. from Carnegie Mellon University and his Ph.D. from the University of Washington in 1997. His current research interests include applications of AI and machine learning to the internet, specifically in the areas of personalized information services; information-extraction, -retrieval, and integration; and intelligent browsing assistants. His e-mail address is nick@ucd.ie.

Marie-Christine Rousset is a professor of computer science at the University of Paris-Sud and a member of the Institut Universitaire de France. After studying mathematics at Ecole Normale Supérieure (Fontenay-aux-Roses) and the University of Paris 6, she received her Ph.D. and These d'Etat in computer science in 1982 and 1988, respectively, at the University of Paris-Sud. Since 1989, she has been leading the Artificial Intelligence and Inference Systems Group in the Computer Science Laboratory at the University of Paris-Sud. Her areas of research are knowledge representation and verification of knowledge bases. Currently, she is jointly leading the PICSEL and GAEL projects that combine knowledge representation and database techniques for information integration and electronic commerce. Her e-mail address is mcr@lri.lri.fr.