INFORMATION BROKERING ACROSS HETEROGENEOUS DIGITAL DATA

A Metadata-based Approach

The Kluwer International Series on
ADVANCES IN DATABASE SYSTEMS

Series Editor

Ahmed K. Elmagarmid

Purdue University
West Lafayette, IN 47907

Other books in the Series:

DATA DISSEMINATION IN WIRELESS COMPUTING ENVIRONMENTS, *Kian-Lee Tan and Beng Chin Ooi*; ISBN: 0-7923-7866-0

MIDDLEWARE NETWORKS: Concept, Design and Deployment of Internet Infrastructure, *Michah Lerner, George Vanecek, Nino Vidovic, Dad Vrsalovic;* ISBN: 0-7923-7840-7

ADVANCED DATABASE INDEXING, *Yannis Manolopoulos, Yannis Theodoridis, Vassilis J. Tsotras;* ISBN: 0-7923-7716-8

MULTILEVEL SECURE TRANSACTION PROCESSING, *Vijay Atluri, Sushil Jajodia, Binto George* ISBN: 0-7923-7702-8

FUZZY LOGIC IN DATA MODELING, *Guoqing Chen* ISBN: 0-7923-8253-6

INTERCONNECTING HETEROGENEOUS INFORMATION SYSTEMS, *Athman Bouguettaya, Boualem Benatallah, Ahmed Elmagarmid* ISBN: 0-7923-8216-1

FOUNDATIONS OF KNOWLEDGE SYSTEMS: With Applications to Databases and Agents, *Gerd Wagner* ISBN: 0-7923-8212-9

DATABASE RECOVERY, *Vijay Kumar, Sang H. Son* ISBN: 0-7923-8192-0

PARALLEL, OBJECT-ORIENTED, AND ACTIVE KNOWLEDGE BASE SYSTEMS, *Ioannis Vlahavas, Nick Bassiliades* ISBN: 0-7923-8117-3

DATA MANAGEMENT FOR MOBILE COMPUTING, *Evaggelia Pitoura, George Samaras* ISBN: 0-7923-8053-3

MINING VERY LARGE DATABASES WITH PARALLEL PROCESSING, *Alex* A. *Freitas, Simon H. Lavington* ISBN: 0-7923-8048-7

INDEXING TECHNIQUES FOR ADVANCED DATABASE SYSTEMS, *Elisa Bertino, Beng Chin Ooi, Ron Sacks-Davis, Kian-Lee Tan, Justin Zobel, Boris Shidlovsky, Barbara Catania* ISBN: 0-7923-9985-4

INDEX DATA STRUCTURES IN OBJECT-ORIENTED DATABASES, *Thomas A. Mueck, Martin L. Polaschek* ISBN: 0-7923-9971-4

DATABASE ISSUES IN GEOGRAPHIC INFORMATION SYSTEMS, *Nabil R. Adam, Aryya Gangopadhyay* ISBN: 0-7923-9924-2

VIDEO DATABASE SYSTEMS: Issues, Products, and Applications, *Ahmed K. Elmagarmid, Haitao Jiang, Abdelsalam A. Helal, Anupam Joshi, Magdy Ahmed* ISBN: 0-7923-9872-6

REPLICATION TECHNIQUES IN DISTRIBUTED SYSTEMS, *Abdelsalam A. Helal, Abdelsalam A. Heddaya, Bharat B. Bhargava* ISBN: 0-7923-9800-9

INFORMATION BROKERING ACROSS HETEROGENEOUS DIGITAL DATA
A Metadata-based Approach

by

VIPUL KASHYAP
Member of Technical Staff
Micro-electronics and Computer Technology Corporation (MCC)
3500, W. Balcones Center Drive
Austin, TX 78759

Current address:
Research Scientist
Applied Research, Telcordia Technologies
MCC-1G332R, 445 South Street
Morristown, NJ 07960

AMIT SHETH
Director, Large Scale Distributed Information Systems Laboratory
Department of Computer Science
The University of Georgia
Athens, GA 30602

KLUWER ACADEMIC PUBLISHERS
New York / Boston / Dordrecht / London / Moscow

eBook ISBN: 0-306-47028-4
Print ISBN: 0-792-37883-0

Created in the United States of America

Visit Kluwer Online at: http://kluweronline.com
and Kluwer's eBookstore at: http://ebooks.kluweronline.com

To my parents, Yogendra and Vibha, whose blessings and encouragement have played an active life-long role in my work, and my wife Nitu for being patient with me.

-Vipul

To the members of the InfoHarness and InfoQuilt team at the LSDIS Lab, who helped us in thinking about metadata and semantics.

-Amit and Vipul

Contents

List of Figures

List of Tables

Preface

Information intermediation is the foundation stone of some of the most successful Internet companies, and is perhaps second only to the Internet Infrastructure companies. On the heels of information integration and interoperability, this book on information brokering discusses the next step in information interoperability and integration.

The emerging internet economy based on burgeoning B2B and B2C trading, will soon demand semantics-based information intermediation for its feasibility and success. Even as we speak, new B2B ventures are involved in the "rationalization" of new vertical markets and construction of domain specific product catalogs. In this book we provide approaches for re-use of existing vocabularies and domain ontologies as a basis for this rationalization and provide a framework based on inter-ontology interoperation. Infrastructural trade-offs that identify optimizations in performance and scalability of web sites will soon give way to information based trade-offs as alternate rationalization schemes come into play, and the necessity of interoperating across these schemes is realized.

The complex issues of intermediation as conceived by information brokering involve the following:

- Construction of information content descriptions that resolve information "impedance" or mismatch between what is required by the information consumer and presented by the information provider.

- Using "information brokering" to provide intermediation to resolve impedance at multiple levels. Architectures for brokering/intermediation should provide a framework for resolution of impedance at different levels, and may be multi-level in nature. New paradigms such as agent based architectures need to be evaluated for their applicability to information intermediation.

- Exploration of the critical role of Metadata. Information content descriptions may be at different levels, such as content-independent, representation dependent, representation independent or domain specific. The descriptions enable resolution of impedances at various levels. Domain specific metadata descriptions are crucial for intermediation based on the semantics (meaning and use) of the information content.
- Using ontology or vocabulary to construct metadata descriptions for aligning the "world views" of consumers and providers. Techniques for re-using existing vocabularies/ontologies and interoperation across them need to be developed to perform this type of intermediation.
- Characterizing loss of information that inevitably arises in semantics-based intermediation due to differences in world views between the consumers and providers. Information-based trade-offs need to be characterized and computed to provide effective semantics-based intermediation.

This book's intended readers are researchers, software architects and CTOs, and advanced product developers dealing with information intermediation issues in the context of e-commerce (B2B and B2C), information technology professionals in various vertical markets (e.g., geo-spatial information, medicine, auto), and practically all librarians at higher-education, technical institutions and universities.

VIPUL KASHYAP AND AMIT SHETH

Acknowledgments

Amit Sheth and I have thoroughly enjoyed intense discussion sessions with each other which we often had in the course of working on this book. Many new research directions and possibilities emerged from those discussions and have played a critical role in enhancing and adding value to this book. This was supplemented by our interactions with members of the InfoHarness and InfoQuilt projects, that helped us in our thinking about metadata and semantics.

Special thanks goes to Eduardo Mena, the heated discussions with whom resulted in the OBSERVER system. Mention needs to be made of the members of the InfoSleuth project team that I worked with. Dr. Marek Rusinkiewicz officially my manager at Micro-electronics and Computer Technology Corporation (MCC), has been a great source of inspiration and help in my research endeavors. I am also indebted to my managers at Telcordia, Sid Dalal and Gardner Patton for their encouragement and advice. Gardner's comments on the draft manuscript were very useful and helped in improving the readability and presentation of content in this book.

Finally, I am grateful to my wife, Nitu, who was gracious enough to donate her time (which we would have otherwise spent together) towards the editing and proof reading of this book.

Foreword

Computing is fast becoming ubiquitous and pervasive. It is *ubiquitous* because computing power and access to the Internet is being made available everywhere; it is pervasive because computing is being embedded in the very fabric of our environment. Xerox Corp. has recently coined the phrase "smart matter" to capture the idea of computations occurring within formerly passive objects and substances. For example, our houses, our furniture, and our clothes will contain computers that will enable our surroundings to adapt to our preferences and needs. New visions of interactivity portend that scientific, commercial, educational, and industrial enterprises will be linked, and human spheres previously untouched by computing and information technology, such as our personal, recreational, and community life, will be affected.

However, when there is information everywhere and all manner of things are interconnected, there arise the problems of information overload and misunderstandings. Dogbert, from Scott Adams' Dilbert, describes the situation as "Information is gushing toward your brain like a fire hose aimed at a teacup". This book analyzes the problems of information overload and misunderstandings and provides a solution: a way for all of the different devices, components, and computers to understand each other, so that they will be able to work together effectively and efficiently. This is a powerful and important advance.

Dr. Vipul Kashyap and Professor Amit Sheth have long been leaders in the area of information system semantics and are widely known as two of the area's most dedicated, productive, and insightful researchers. Together in this book they have crafted a coherent vision of the single most important element of a distributed heterogeneous information system: the information broker. Previously, Kashyap has been the architect of a broker-based multiagent system for cooperative information access. Sheth has been an innovator of metadata-based approaches to the integration of heterogeneous semantics for database systems and workflow systems. Together, their expertise is complementary and forms the unique perspective of this book.

The essential agent-based architecture that they describe and analyze is becoming canonical. Agents are used to represent users, resources, middleware, security, execution engines, ontologies, and brokering. As the technology advances, we can expect such specialized agents to be used as standardized building blocks for information systems. Two trends lend credence to such a prediction.

First, software systems in general are being constructed with larger components, such as ActiveX and JavaBeans, which are becoming closer to being agents themselves. They have more functionality than simple objects, respond to events autonomously, and, most importantly, respond to system builders at development-time, as well as to events at run-time.

Second, there is a move toward more cooperative information systems, in which the architecture itself plays an important role in the effectiveness of the system, as opposed to traditional software systems where effectiveness depends on the quality of the individual components. These are the architectures of standardized agents that Kashyap and Sheth elucidate. Architectures based on standardized agent types should be easier to develop, understand, and use. Perhaps most important of all, these architectures will make it easier for separately developed information systems to interoperate.

Among the reasons why agents are attractive, there are two main ones of interest here. One, agents enable the construction of modular systems from heterogeneous pieces, potentially created by any number of vendors. Two, the agents themselves embody diverse knowledge, reasoning approaches, and perspectives. This diversity is sometimes essential, because the agents represent people or business interests that have different goals and motivations. Diversity can sometimes be added in by design: it can make an agent system more robust by enabling a variety of viewpoints to be represented and exploited.

However, agents are typically complex pieces of software, so the question arises whether a set of different agents would unnecessarily add to a system's complexity. The more kinds of agents there are, the harder it might be to build and maintain them. Fortunately, this turns out to be a false concern. The agents have to be diverse in content, e.g., knowledge, reasoning techniques, and interaction protocols, but not in the form in which that content is realized, e.g., the language or toolkit with which they are constructed. Problems arise through unnecessary heterogeneity in construction; the cost of necessary heterogeneity in content is more than recovered through the flexibility it offers.

In summary, the results in this book are applicable not only to the huge amount of information available globally over the World-Wide Web, but also to the diverse information soon to be available locally over household, automobile, and environment networks. I am excited by the possibilities for new applications and uses for information that are engendered by this book, as well

as by the challenges that remain. This book provides a solid foundation for advances in distributed heterogeneous information systems.

Professor Michael N. Huhns,
Director, Center for Information Technology,
University of South Carolina, Columbus, SC