FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

INTEGRATION AND QUERYING OF HETEROGENEOUS, AUTONOMOUS, DISTRIBUTED DATABASE SYSTEMS

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Rukshan Indika Athauda

2000

To: Dean Arthur W. Herriott
College of Arts and Sciences

This dissertation, written by Rukshan Indika Athauda, and entitled Integration and Querying of Heterogeneous, Autonomous, Distributed Database Systems, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

_____
Jun Yuan

_____
Nagarajan Prabhakar

_____
Subbarao Wunnava

_____
Naphtali Rishe, Major Professor

Date of Defense: July 5, 2000

The dissertation of Rukshan Indika Athauda is approved.

_____
Dean Arthur W. Herriott
College of Arts and Sciences

_____
Dean Richard L. Campbell
Division of Graduate Studies

Florida International University, 2000

DEDICATION

I dedicate this dissertation to my parents. I am forever indebt for their guidance, patience, understanding, support and love throughout my whole life.

ACKNOWLEDGMENTS

I wish to thank Dr. Naphtali Rishe, my major professor, who provided me with guidance, direction and support to perform my doctoral research. I wish to thank my committee members Dr. Prabhakaran, Dr. Wunnava and Dr. Yuan for taking the time to review my thesis and for their helpful comments. I would like to thank Dr. Yuan for his support throughout the design and implementation of the project. Dr. Shu-Ching Chen deserves a special note of thanks for providing me with very helpful comments, revisions in paper writing and my dissertation work. Also, my colleagues, Xiaoling Lu's and Xiaobin Ma's efforts in implementing the wrapper project is greatly appreciated. I would like to thank the secretaries, Theresa O'Connel and Maria Monteagudo, who were always willing to help me. I would like to thank, Catherine Hernandez and support staff for promptly responding to our requests. A special note of thanks is extended to the library and its staff for providing access to many research resources without which this project may not be feasible. I deeply regret the fact that I am unable to acknowledge everyone who supported me throughout the years at FIU (especially the excellent faculty). I would like to convey my heartfelt appreciation and gratitude to them.

ABSTRACT OF THE DISSERTATION

INTEGRATION AND QUERYING OF HETEROGENEOUS, AUTONOMOUS, DISTRIBUTED DATABASE SYSTEMS

by

Rukshan Indika Athauda

Florida International University, 2000

Miami, Florida

Professor Naphtali Rishe, Major Professor

Today, databases have become an integral part of information systems. In the past two decades, we have seen different database systems being developed independently and used in different applications domains. Today's interconnected networks and advanced applications, such as data warehousing, data mining & knowledge discovery and intelligent data access to information on the Web, have created a need for integrated access to such heterogeneous, autonomous, distributed database systems. Heterogeneous/multidatabase research has focused on this issue resulting in many different approaches. However, a single, generally accepted methodology in academia or industry has not emerged providing ubiquitous intelligent data access from heterogeneous, autonomous, distributed information sources.

This thesis describes a heterogeneous database system being developed at High-performance Database Research Center (HPDRC). A major impediment to ubiquitous

deployment of multidatabase technology is the difficulty in resolving semantic heterogeneity. That is, identifying related information sources for integration and querying purposes. Our approach considers the semantics of the meta-data constructs in resolving this issue. The major contributions of the thesis work include: (i.) providing a scalable, easy-to-implement architecture for developing a heterogeneous multidatabase system, utilizing Semantic Binary Object-oriented Data Model (Sem-ODM) and Semantic SQL query language to capture the semantics of the data sources being integrated and to provide an easy-to-use query facility; (ii.) a methodology for semantic heterogeneity resolution by investigating into the extents of the meta-data constructs of component schemas. This methodology is shown to be correct, complete and unambiguous; (iii.) a semi-automated technique for identifying semantic relations, which is the basis of semantic knowledge for integration and querying, using shared ontologies for context-mediation; (iv.) resolutions for schematic conflicts and a language for defining global views from a set of component Sem-ODM schemas; (v.) design of a knowledge base for storing and manipulating meta-data and knowledge acquired during the integration process. This knowledge base acts as the interface between integration and query processing modules; (vi.) techniques for Semantic SQL query processing and optimization based on semantic knowledge in a heterogeneous database environment; and (vii.) a framework for intelligent computing and communication on the Internet applying the concepts of our work.

TABLE OF CONTENTS

LIST OF FIGURES