

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

© Copyright 2000 by High-performance Database Research Center at

Florida International University

All rights reserved.

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 Rische, 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 Rische, 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

CHAPTER	PAGE
1. INTRODUCTION	1
1.1 Related Work	5
1.2 Our Work	7
1.2.1 Contributions of Thesis	8
1.2.2 Limitations	11
1.2.3 Outline of Thesis	11
2. HETEROGENEOUS DISTRIBUTED DATABASE SYSTEM	13
2.1 Related Work	13
2.2 Our Work	16
2.2.1 Semantic Binary Object-oriented Data Model	16
2.2.2 Semantic SQL Query Language	19
2.2.3 System Architecture	25
2.3 Benefits	29
3. SEMANTIC HETEROGENEITY RESOLUTION	34
3.1 Related Work	36
3.2 Our Work	41
3.2.1 Foundations of Semantic Knowledge	42
3.2.1.1 Semantic Relations	43
3.2.1.2 Object Equivalence	46
3.2.1.3 Boundary Conditions	50
3.2.2 Identification of Semantic Relations	52
3.2.2.1 Relevant Work	53
3.2.2.1.1 Ontological Foundations	53
3.2.2.1.2 Classification Techniques	58
3.2.2.2 Methodology	60
3.2.2.2.1 Step 1: Conversion to Sem-ODM	61
3.2.2.2.2 Step 2: Obtaining Property Functions	63
3.2.2.2.3 Step 3: Mapping to Shared Ontology	66
3.2.2.2.4 Step 4: Discovering Semantic Relations	69
3.3 Summary	72
4. SCHEMATIC HETEROGENEITY RESOLUTION	74
4.1 Related Work	74
4.2 Our Work	75
4.2.1 SemOSQL/M	76
4.2.2 Schema-level Conflicts and Resolutions	78
4.2.2.1 Naming Conflicts	79
4.2.2.2 Data Conflicts	80
4.2.2.3 Attribute Type Conflicts	82

4.2.2.4	Attribute Granularity Conflicts	83
4.2.2.5	Missing Attribute Conflicts	84
4.2.2.6	Missing Attributes With Implicit Values	86
4.2.2.7	Basic Relations	87
4.2.2.8	Composite Relations	89
4.2.2.9	Inter-schema Relations	91
4.2.2.10	Category Inclusion Conflicts	92
4.2.2.11	Attribute Inclusion Conflicts	93
4.2.2.12	Category-versus-Attribute Conflicts	94
4.2.3	Handling Inconsistent Data	97
4.2.4	Knowledge Management in Database Integration	98
4.2.4.1	Knowledge Base	99
4.2.4.1.1	Knowledge Bases at Component Sites	99
4.2.4.1.2	Knowledge Base at Global Site	106
4.2.4.2	A Tool used for Global View Definition	110
4.3	Summary	113
5	QUERY PROCESSING	117
5.1	Related Work	118
5.2	Our Work	121
5.2.1	Step 1: Scanning, Parsing and Semantic Checking	122
5.2.2	Step 2: Relational Algebra and Logical Optimization	126
5.2.3	Step 3: Expanding the Virtual Tables	129
5.2.4	Step 4: Global Query Optimization	131
5.2.5	Step 5: Generating Subqueries	141
5.3	Summary	142
6	A FRAMEWORK FOR THE INTELLIGENT WEB	144
6.1	Related Work	146
6.2	Our Work	149
6.2.1	Framework for the Internet	150
6.2.2	A Futuristic View of E-Commerce Applications	154
6.2.3	Future Research Issues and Technology Directions	157
6.3	Summary	158
7	CONCLUSION	160
	LIST OF REFERENCES	163
	APPENDICES	175
	VITA	189

LIST OF FIGURES

FIGURE	PAGE
1. Schema of a Relational Database Developed for a University Application	1
2. Semantic Database Schema in Computer Science Department of University Consisting Information of Students and Projects	2
3. Relational Schema Equivalent to the Sem-ODM Schema Presented in Figure 2 .	17
4. (a.) Semantic SQL Query posed on the Sem-ODM Schema (b.) SQL Query posed on the Equivalent Relational Schema	23
5. Architecture of Heterogeneous Distributed Database System	24
6. Semantic Schema of a University Application	34
7. (a.) Schema of Database DB_1 in Administration Office of Company A (b.) Schema of Database DB_2 in Lab L of Company A	38
8. Integrated Schema for Schemas Presented in Figure 7	39
9. All Possible Scenarios for $EXT(A)$ and $EXT(B)$: (a.) $EXT(A) = EXT(B)$ (b.) $EXT(A) \subseteq EXT(B)$ (c.) $EXT(A) \cap EXT(B) \neq \phi$ (d.) $EXT(A) \cap EXT(B) = \phi$	44
10. (a.) Category of Database DB_1 Containing Information of Students in University A (b.) Category of Database DB_2 Containing Information of Students in University A	47
11. A Semantic Net with ISA and AKO Links	59
12. Schema in Sem-ODM	63
13. Schema for Database DB_2 in Example 1	67
14. Schemas with Naming Conflicts	79
15. Schemas with Data Conflicts	81

16. Schemas with Missing Attribute Conflicts	85
17. Schemas having Missing Attributes with Implicit Value Conflicts	86
18. Schemas with Basic Relations	88
19. Schemas with Composite Relations	90
20. Schemas with Inter-schema Relations	91
21. Schemas with Category Inclusion Conflicts	93
22. Schemas with Category-versus-Attribute Conflicts	95
23. Semantic Schema Created By Transforming a Relational Schema	104
24. Sem-ODM Schema of DB ₂	108
25. Tree Representation of a Virtual Table	130
26. Query Execution Plan Without Considering Semantic Relations	135
27. Query Execution Plan Without Considering Semantic Relations	140
28. Query Execution Plan Considering Semantic Relations	140
29. Overall View of Proposed Framework for the Internet	151
30. A Model for E-Commerce Applications on the Web	152