

การรวมข้อมูลสารสนเทศจากแหล่งสารสนเทศวิจิตรพันธุโดยวิธีออนไลน์

นางงามนิจ อาจอินทร์



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2546

ISBN 974-17-3952-4

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

I 2150 77 1 5

ONTOLOGY-BASED APPROACH FOR GATHERING
THE HETEROGENEOUS INFORMATION SOURCES

Mrs. Ngamnij Arch-int

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Computer Science

Department of Mathematics

Faculty of Science

Chulalongkorn University

Academic Year 2003

ISBN 974-17-3952-4

Thesis Title Ontology-based Approach for Gathering the Heterogeneous
Information Sources
By Mrs. Ngamnij Arch-int
Field of Study Computer Science
Thesis Advisor Assistant Professor Peraphon Sophatsathit, Ph.D.

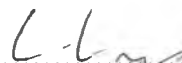
Accepted by the Faculty of Science, Chulalongkorn University in Partial
Fulfillment of the Requirements for the Doctor's Degree

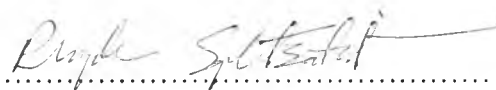



..... Dean of Faculty of Science

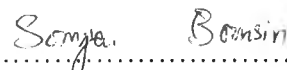
(Professor Piamsak Menasveta, Ph.D.)


THESIS COMMITTEE


..... Chairman
(Professor Chidchanok Lursinsap, Ph.D.)


..... Thesis Advisor
(Assistant Professor Peraphon Sophatsathit, Ph.D.)


..... Member
(Assistant Professor Pattarasinee Bhattarakosol, Ph.D.)


..... Member
(Assistant Professor Somjai Boonsiri, Ph.D.)


..... Member
(Surapant Meknavin, Ph.D.)

งามนิจ อาจอินทร์ : การรวมข้อมูลสารสนเทศจากแหล่งสารสนเทศวิวิธพันธุ์โดยวิธีออนโทโลยี (Ontology-based Approach for Gathering the Heterogeneous Information Sources)

อ. ที่ปรึกษา: ผศ. ดร. พีระพนธ์ โสพัศสถิตย์, จำนวนหน้า 144 หน้า. ISBN 974-17-3952-4

การเข้าถึงและการรวบรวมข้อมูลสารสนเทศจากแหล่งสารสนเทศวิวิธพันธุ์ เป็นปัญหาหลักของระบบการประมวลผลแบบกระจาย ซึ่งสาระสำคัญของปัญหานั้นคือความหลากหลายทางด้านความหมายของข้อมูล หลายระบบถูกนำเสนอเพื่อแก้ปัญหาที่เกิดขึ้น ตั้งแต่ระบบตัวกลาง (Mediator-based Systems) ไปจนถึงระบบวงจรรพรรณนา (Description logic-based systems) อย่างไรก็ตาม วิธีการแก้ปัญหานี้ในปัจจุบัน ยังมีข้อจำกัดทางด้านความยืดหยุ่น ความสามารถในการขยายระบบ การปฏิบัติงานร่วมกัน และความแข็งแกร่งของระบบ

วิทยานิพนธ์นี้ได้นำเสนอ ระบบการรวบรวมข้อมูลสารสนเทศแบบสื่อความหมาย (Semantic Information Gathering Approach) ขึ้น สำหรับการเข้าถึงและการรวบรวมข้อมูลสารสนเทศ จากแหล่งสารสนเทศวิวิธพันธุ์บนเครือข่ายอินเทอร์เน็ต สถาปัตยกรรมของระบบออกแบบตามหลักการของโครงสร้างที่แบ่งระดับชั้นการทำงานเป็นอิสระจากกัน (Layered-architecture) โดยการผนวกซอฟต์แวร์ตัวแทนที่เคลื่อนที่ได้ (Mobile agent) ทำหน้าที่เป็นตัวเชื่อมระหว่าง ลูกข่ายและแม่ข่าย (Client/Server) เพื่อลดภาระงานที่หนักอึ้งเนื่องจากสถาปัตยกรรม ลูกข่าย/แม่ข่าย วิทยานิพนธ์นี้ยังนำเสนอ พจนานุกรมข้อมูล (Metadata Dictionary) ซึ่งเป็นองค์ประกอบที่สำคัญของสถาปัตยกรรมข้างต้น เพื่อแก้ปัญหาคความหลากหลายทางด้านความหมายของข้อมูล องค์ประกอบของพจนานุกรมข้อมูลนี้ ได้จากออนโทโลยีเฉพาะกิจ (Domain Ontology) ซึ่งนิยามจากหลักการเชิงวัตถุและทฤษฎีเซต เพื่อให้ระบบสามารถปฏิบัติงานร่วมกันได้ดี อันเป็นการประยุกต์ที่เหมาะสมกับการทำงานบนอินเทอร์เน็ต ในส่วนของระบบใช้ภาษา XML เป็นภาษาสำหรับการแสดงพจนานุกรมข้อมูล วิทยานิพนธ์นี้ยังครอบคลุมรายละเอียดของการทำงานทั้งหมด สถาปัตยกรรมของซอฟต์แวร์ตัวแทน กระบวนการสร้างพจนานุกรมข้อมูลบนพื้นฐานของออนโทโลยี องค์ประกอบและการแสดงองค์ประกอบของพจนานุกรม รวมถึงกระบวนการค้นหาข้อมูลจากแหล่งสารสนเทศวิวิธพันธุ์ด้วยพจนานุกรมข้อมูล

ภาควิชา คณิตศาสตร์
สาขาวิชา วิทยาการคอมพิวเตอร์
ปีการศึกษา 2546

ลายมือชื่อนิสิต ... Ngamniy Arch-inl-
ลายมือชื่ออาจารย์ที่ปรึกษา ... Pich Sritat

4373810423: MAJOR COMPUTER SCIENCE

KEYWORDS: HETEROGENEOUS INFORMATION SOURCES/ DOMAIN ONTOLOGY/ XML-BASED METADATA DICTIONARY/ QUERY PROCESSING

NGAMNIJ ARCH-INT: ONTOLOGY-BASED APPROACH FOR GATHERING THE HETEROGENEOUS INFORMATION SOURCES. THESIS ADVISOR: ASST. PROF. PERAPHON SOPHATSATHIT, Ph.D., 144 pp. ISBN 974-17-3952-4.

The problem of accessing and integrating heterogeneous information sources has become the center stage in the distributed processing environment. One of the important issues stemming from accessing these heterogeneous sources is semantic heterogeneity. A number of systems have been proposed to address this issue, ranging from mediator-based systems to description logic-based systems. However, the current methodologies for accessing heterogeneous sources offer limited flexibility, scalability, interoperability, and robustness.

This dissertation proposes the Semantic Information Gathering Approach, hereafter *SIGA*, a system for accessing and integrating heterogeneous sources on the WWW. The reference architecture is based on layered-architecture incorporating mobile agents as the client/server connectivity to eliminate traditional client/server overhead. A metadata dictionary which is an essential component of the reference architecture is also proposed for solving semantic heterogeneity. The metadata dictionary is derived from domain ontology where the constituent components are defined in terms of object-oriented principle and set theory. To enable system-wide interoperability suitable for a Web-based environment, the XML technology is selected as the language for expressing the metadata dictionary contents. Consequently, this dissertation also covers the overall system and agent architecture, the modeling process of the ontology-based metadata dictionary, the components and representation of the metadata dictionary, and the querying process in accessing and integrating the heterogeneous sources through the metadata dictionary.

Department of Mathematics

Field of study Computer Science

Academic year 2003

Student's signature.....*Ngamni J. Arch-int.*

Advisor's signature.....*Peraphon Sophatsathit*



ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Dr. Peraphon Sophatsathit, for his invaluable guidance, constructive comments, and continuous encouragement throughout the period of this study. Gratitude is also extended to Prof. Chidchanok Lursinsap, Dr. Pattarasinee Bhattarakosol, Dr. Somjai Boonsiri, and Dr. Surapant Meknavin, my examination committee members, for their reading, comments and suggestions in making this dissertation a reality.

I am very grateful to Dr. Yuefeng Li, and Dr. Paul Roe, who provided me the opportunity as a visiting scholar at the School of Software Engineering and Data Communications, Queensland University of Technology, Brisbane, Australia, as well as their suggestions and assistance during my visiting for 8 months.

I also would like to express my gratitude to the Ministry of University Affairs for all financial support throughout my study.

A special thank is given to the members of AVIC and staff for their help and friendship.

Finally, my deepest gratitude goes to my beloved father, mother, husband, daughter and son, for their invaluable love, support, and encouragement throughout my whole life.

TABLE OF CONTENTS

	page
ABSTRACT IN THAI	iv
ABSTRACT IN ENGLISH	v
ACKNOWLEDGMENTS	vi
TABLE OF CONTENTS	vii
LIST OF FIGURES	xi
LIST OF TABLES	xiii
CHAPTER 1 INTRODUCTION	1
1.1 The Objectives	4
1.2 Procedure and Outline	4
1.3 Benefits of the Dissertation	5
CHAPTER 2 THEORETICAL BACKGROUND	7
2.1 Agent Technology	7
2.1.1 Agent Definitions	7
2.1.2 Agent Model	8
2.1.3 Type of Agents	10
2.1.4 Mobile Agents	10
2.1.5 The Benefits of Mobile Agent	11
2.2 Ontology-based Approach	13
2.2.1 The Components of Ontologies	13
2.2.2 Type of Ontologies	14
2.2.3 Ontology Representation	14
2.2.4 Ontology Architectures	15
2.2.5 Applications of Ontologies	17
2.3 XML Technology	18
2.4 Information Integration Architectures	20
2.4.1 Data Warehouse Architecture	21

TABLE OF CONTENTS (Cont'd)

2.4.2 Mediator-Wrapper Architecture	22
2.5 The Information Source Connectivity	27
CHAPTER 3 SIGA: SEMANTIC INFORMATION GATHERING APPROACH	29
3.1 An Overview of the SIGA Reference Architecture.....	29
3.2 Presentation Layer	32
3.3 Mediator Layer	32
3.3.1 User Interface Agent	32
3.3.2 Managing Agent	33
3.3.3 Metadata Dictionary	35
3.4 Search Layer	36
3.5 Resource Layer	36
CHAPTER 4 ONTOLOGY-BASED METADATA DICTIONARY MODELING.	38
4.1 Ontology-based Metadata Dictionary Modeling	38
4.1.1 Schema Translation	39
4.1.2 Schema Restructuring	39
4.1.3 Schema Integration	41
4.1.4 Ontology Extraction	42
4.2 Metadata Dictionary Management	43
CHAPTER 5 ONTOLOGY-BASED METADATA DICTIONARY	
COMPONENTS.....	47
5.1 Ontology-based Metadata Dictionary Components	47
5.1.1 Virtual Concept	47
5.1.2 Relationships	50
5.1.3 Physical Source Configurations	53
CHAPTER 6 XML-BASED METADATA DICTIONARY.....	54
6.1 Structural Design of XML-DTD from Domain Ontology Components.....	54
6.1.1 The Conceptual Level of Design Abstraction	54
6.1.2 The Physical Level of Design Abstraction	57

TABLE OF CONTENTS (Cont'd)

6.2 XML-DTD Metadata Dictionary Structure	58
6.3 Construction Rules	59
CHAPTER 7 THE MODELING OF THE ONTOLOGY-BASED METADATA	
DICTIONARY: A CASE STUDY	62
7.1 An Example of the Semantic Heterogeneity	62
7.2 Domain Ontology Representation	64
7.2.1 The Conceptual Level Representation	64
7.2.2 The Physical Level Representation	65
7.3 The XML-based Metadata Dictionary Representation	66
CHAPTER 8 QUERY PROCESSING FOR THE HETEROGENEOUS	
INFORMATION SOURCES USING METADATA DICTIONARY APPROACH 69	
8.1 The Accessing Process of the Heterogeneous Information Sources	70
8.2 The Integrating Process of the Heterogeneous Information Sources	79
8.2.1 Single Source Integration	80
8.2.2 Multiple Sources Integration	82
8.3 The Query Validation	90
8.3.1 The Query Requirement Correctness Validation	90
8.3.2 The Result Correctness Validation	90
CHAPTER 9 LITERATURE REVIEW AND COMPARISON..... 91	
9.1 The Mediator-based Systems	91
9.1.1 TSIMMIS	92
9.2 The Description Logic-based Systems	94
9.2.1 Information Manifold	95
9.2.2 OBSERVER	97
9.3 Comparative Characterization with SIGA	99

TABLE OF CONTENTS (Cont'd)

CHAPTER 10 CONCLUSIONS AND RECOMMENDATIONS	103
10.1 Conclusions	103
10.2 Recommendations	104
REFERENCES	105
APPENDICES	113
Appendix A: The XML-DTD of the Metadata dictionary	114
Appendix B: Examples of valid XML Document conforming to the proposed XML-DTD	116
Appendix C: Algorithms for mapping virtual schema to physical schema of the metadata dictionary	123
Appendix D: Algorithms for transforming the XML results returned from a single source into a unified XML-based data	126
Appendix E: Algorithms for multiple source integration using merging technique	129
Appendix F: Algorithms for multiple source integration using join technique	133
Appendix G: Algorithms for verifying result correctness.....	136
Appendix H: Case study: Implementation and installation guide of SIGA	138
BIOGRAPHY	144

LIST OF FIGURES

	page
2.1 Agent model	9
2.2 The ontology architectures	17
2.3 Examples of HTML and XML documents	19
2.4 Database access from a Web browser	21
2.5 A data warehouse architecture	22
2.6 A mediator-wrapper architecture	23
2.7 Static integration architecture.....	25
2.8 Dynamic integration architecture	26
2.9 Hooking database system and the Web	27
3.1 The reference architecture of SIGA	31
3.2 The user interface agent architecture.....	33
3.3 The managing agent architecture	34
3.4 The internal process of result integrator module	35
3.5 The resource agent architecture	37
4.1 Extraction of the ontology-based metadata dictionary by domain ontology modeling	39
4.2 Atomic conformation principles	40
4.3 The relationship between ontology-based metadata dictionary and the underlying physical schemas.....	41
4.4 Two levels of domain ontology extracted from a global conceptual schema ..	42
4.5 The metadata dictionary components	45
5.1 The subsumption hierarchy of the ontology	51
5.2 The IS-PART-OF relationship	52
6.1 The XML-DTD structure at the conceptual level of design abstraction	55
6.2 The XML-DTD structure at the physical level of design abstraction	58
6.3 The XML-DTD metadata dictionary structure.....	59
7.1 An overview of the hierarchical concepts of the university system	63
7.2 Three different data models of physical information sources	64

LIST OF FIGURES (Cont'd)

7.3	The logical ontology structure at the conceptual level of abstraction	65
7.4	A portion of internal structure of the ontology at the physical level of abstraction	66
7.5	A portion of the XML document structure conforming to earlier XML-DTD.	67
7.6	A portion of the XML document based on metadata dictionary	68
8.1	An example of the metadata dictionary contents represented by a labeled tree	71
8.2	An example of the global transaction simplification.....	72
8.3	A accessing process of the heterogeneous information sources.....	73
8.4	Two initial sub-transactions generated from the substitution process	76
8.5	A portion of metadata dictionary illustrating the replicated data	77
8.6	The XML returned results from Source2 to be sent to the managing agent.....	81
8.7	The unified XML-based data generated from the managing agent	81
8.8	Multiple sources integration by merging the XML documents into the unified XML document	85
8.9	Multiple sources integration by merging the XML-DTD of each source into the unified XML-DTD	86
8.10	An example of the global transaction decomposition into sub-transactions.....	87
8.11	Multiple sources integration by joining the XML documents into the unified XML document.....	88
8.12	Multiple sources integration by joining the XML-DTD of each source into the unified XML-DTD	89
9.1	The mediator-based information systems architecture	92
9.2	The TSIMMIS architecture	93
9.3	Examples of the OEM objects	94
9.4	Examples of source description related to the world view in Table 9.1	95
9.5	The information manifold architecture	96
9.6	The OBSERVER: An architecture to support query processing	98

LIST OF TABLES

	page
9.1 Examples of a class hierarchy representing world view	95
9.2 Comparison of various ontology systems and SIGA characteristics.....	100