SEMANTIC TRANSFORMATIONS FOR XML QUERIES

By

Dung Xuan Thi LE
BSc. (*Swinburne*), *MSc.* (*La Trobe*)

THIS THESIS IS PRESENTED FOR

THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTING, FACULTY OF SCIENCE

2012

©Dung Xuan Thi Le, 2012

Statement of Authorship

I certify that the work in this thesis titled “Semantic Transformations for XML Queries” has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree to any other university or institution other than Macquarie University.

I also certify that the thesis is an original piece of research and it has been written by me. Any help and assistance I have received in my research work and the preparation of thesis itself have been appropriately acknowledged.

In addition, I certify that all information resources and literature used are indicated in the thesis.

Dung Xuan Thi LE

Acknowledgements

I am deeply indebted to Prof. Mehmet Orgun for his supervision, support and advice which have been a great inspiration. The completion of this thesis would not have been possible without his patience, input and evaluation of my earlier efforts. I would also like to thank my co-supervisor Dr. Peter Busch who has also evaluated my drafts. I would like to extend my acknowledgement and thanks to my former supervisors Dr. Wenny Rahayu, Dr Eric Pardede and Dr. Stephane Bressan. I especially wish to acknowledge the efforts of Dr. Eric Pardede; this thesis would not have been possible without his dedicated time to evaluate my draft, friendship, support and encouragement. I thank Dr Michael Hill for his time to check and proofread part of my work.

I would also like to thank Dr. Saqib Ali, Oldooz Dianat, Dr. Paul Conilione, Dr. Regina Ryan, Lachlan Williams, Mary Law, Trairat Em-Oj and Elizabeth Le for their friendship, help and support.

I am grateful to the families of Dinh, Kelly, Bob and Bill and to the Vincent De Paul Society for their love and generosity. Without their kindness, I would not be here in Australia today.

I would also like to thank the directors of GMS and MMM Technology, my much earlier employers, who gave me many opportunities to gain invaluable experience during my period of employment. In particular, I thank the Ang family for their kindness, generosity and encouragement of my further study.

This thesis is specially dedicated to a very special person who has had a great influence on what I do and who I have become. His constant support, encouragement and love have made me believe in a better tomorrow.

Saving the most important for last, my heartfelt thanks go to my parents for their unconditional love and sacrifices which have inspired me to make them proud. I thank my dear brothers and sisters for their continuous support and love.

List of External Publications & Statement of Contributions

This thesis contains material that has been published and prepared for submission as follows:

1. *Le, D. X. T., Stéphane Bressan, David Taniar, J. Wenny Rahayu, Eric Pardede: Using semantics for XPath query transformation. IJWGS 6(1):58-94 (2010).*
2. *Le, D. X. T., Bressan, S., Taniar, D., Rahayu, W., "Semantic XPath Query Transformation: Opportunities and Performance". In the 12th International Conference on Database Systems for Advanced Applications (DASFAA), 2007, pp. 994-1000. (Ranked in A category in ERA-2010).*
3. *Le, D. X. T., Bressan, S., Pardede, E., Rahayu, W., Taniar, D., "A Utilization of Schema Constraints to Transform Predicates in XPath Query". In the 21st International Conference on Database and Expert Systems Applications (DEXA), 2010, pp. 331-339. (Ranked in B category in ERA-2010)*

The first paper is the consolidation and revision of work reported in the second and third papers; they resulted from my PhD research at La Trobe University. The co-authors J. Wenny Rahayu and Eric Pardede were my PhD supervisors there. Stéphane Bressan and David Taniar were my external PhD advisors. I collaborated with Stéphane Bressan on background of semantic transformations and experimental design. I collaborated with David Taniar on the area of result presentations. My contributions to these three papers as a whole are 80%.

4. *Le, D. X. T., Bressan, S., Pardede, E., Rahayu, W., Taniar, D., "Semantic Transformation Approach with Schema Constraints for XPath Query Axes". In the 11th International Conference on Web Information System Engineering (WISE), 2010, pp. 456-470. (Ranked in A category in ERA-2010).*

The fourth paper above resulted from my PhD research at La Trobe University. The co-authors J. Wenny Rahayu and Eric Pardede were my PhD supervisors there. Stéphane Bressan and David Taniar were my external PhD advisors with whom I collaborated on the analysis of the experimental result of my research. My contributions to this paper as a whole are 85%.

5. *Le, D. X. T., Pardede, E., "On Using Semantic Transformation Algorithms for XML Safe Update". In the 8th International Conference on Information Systems Technology and its Applications (UNISCON), 2009, 276-278. (Ranked in B category in ERA-2010).*
6. *Le, D. X. T., Pardede, E., "Towards Performance Efficiency in Safe XML Update". In the 8th International Conference on Web Information Systems Engineering (WISE), 2007, pp. 563-572. (Ranked in A category in ERA-2010).*

The fifth and sixth papers above resulted from the research collaboration that combines my research on *semantic transformations* and the *safe update* research conducted by Dr Eric Pardede to propose semantic safe updates for XML queries. My contributions to these papers including the proposed techniques, implementation and analysis of the results as a whole are 65%.

7. *Le, D. X. T., Orgun, M., "Evaluation of Semantic XPath Queries for XML-Enabled Databases". (In Preparation for Journal Submission).*

The seventh paper above is prepared for a journal submission. It has resulted from the work reported in chapters 8 and 9 of this thesis. My contribution to this paper is 80%.

Abstract

The ever-increasing adoption of XML has created a need to ensure that XML query languages perform efficiently. *Query optimization* and *trans*formation for XML query languages, both syntactically and semantically, have received much attention from research communities in recent years. However, due to the fast progress of the application of XML data management solutions, XML-Enabled Database Management Systems still face several challenges. Among these challenges is query processing, especially the processing of XML queries specified with XPath axes and redundancies that may exist in predicates used in XML queries.

Semantic query optimization utilizes constraints in XML schemas to directly optimize a given query with a set of optimization rules. Due to the current complexity of the XML data structure which is enabled by rich semantics in XML Schemas, semantic query optimization should be performed in a more systematic manner. For a complete solution, this research proposes a series of semantic transformations to transform given XML queries to semantically equivalent, but more efficient, XML queries for optimization purposes, by using the semantics provided in XML Schemas. The proposed semantic transformations are grouped into three categories: (1) *Semantic Path Transformations*, (2) *Semantic Transformations for XPath Queries Specified with Predicates,* and (3) *Semantic Transformations for XPath Queries Specified with XPath Axes*.

After a semantic transformation is applied to an XML query, the equivalent semantic XML query can be processed more efficiently by an XML data management system and returns the same result set.

The proposed semantic transformations are then translated into a series of algorithms which are implemented and empirically evaluated for their efficiency and effectiveness. The experimental studies were carried out by using both real data sets (DBLP) and Benchmark data sets (Michigan) to illustrate that the majority of semantic transformations achieved significantly improved performance in XML

query processing; this also enabled the research presented here to identify semantic transformations as optimization devices.

Table of Contents

Statement of Authorship iii

Acknowlegement v

List of External Publications & Statement of Contributions vii

Abstract ix

Table of Contents xi

List of Figures xv

Chapter 1 Introduction 1

1.1 XML: Background 2

1.2 XML Databases 3

1.3 XML Schema & Its Importance 5

1.4 XML: Query Languages 6

1.5 Motivation 7

1.6 Objectives 9

1.7 Scope and Plan of the Thesis 10

Chapter 2 Literature Review 15

2.1 Clarification & Classification of Techniques 15

2.2 Legacy Databases 16

2.3 XML Databases 21

2.3.1 XML Queries without Predicates 22

2.3.2 XML Query with Predicates 31

2.4 Summary and Open Challenges 36

Chapter 3 Problem Definitions 39

3.1 An Overview of Problem Definition 39

3.2 XML and Query Essential Background 42

3.2.1 XML Query Component and Structure 43

3.2.2 Notions of XML Query Structural Processing & Equivalence 49

3.3 Overview of Semantics & Features in XML Schemas 50

3.3.1 Structural Constraints in XML Schema 50

3.3.2 Constraints of Elements in XML Schema 54
3.4 Semantic Transformations – Summary of Problem Definition 56
3.5 Experiment & Performance Evaluation 58
3.5.1 Experimental Design 59
3.5.2 Performance Evaluation 61
3.6 Summary 62
Chapter 4 Derivation of Semantic & Semantic Path Transformation 63
4.1 Derivation of Semantics 63
4.1.1 Deriving Unique Paths 67
4.1.2 Deriving Constraints of Elements 69
4.2 Semantic Path Transformation 70
4.2.1 Semantic Path Expansion 71
4.2.2 Semantic Path Contraction 75
4.2.3 Semantics Conflict Complement 79
4.2.4 Semantics Conflict Detection 83
4.3 Summary 85
Chapter 5 Semantic Transformations for XPath Queries specified with XPath Axis 87
5.1 Query Transformation Direction and Defining Parameters 88
5.1.1 Transformation Direction for XPath Query 89
5.1.2 Defining Global Parameters 90
5.2 Semantic Transformation for Following- or Preceding-sibling Axis 91
5.3 Semantic Transformation for Following or Preceding Axis 98
5.4 Semantic Transformation for Ancestor or Ancestor-or-self Axis 104
5.5 Semantic Transformation for Parent Axis 108
5.6 Semantic Transformation for Descendant or Descendant-or-self Axis 112
5.7 Semantic Transformation for Self Axis 117
5.8 Semantic Transformation for Child Axis 120
5.9 Conclusion 123
Chapter 6 Semantic Transformations for XPath Queries Specified with Predicates 125
6.1 Semantic Transformation for Predicates: Motivation 126
6.2 XPath Query Predicate Structures and Global Parameters 127

6.2.1 XPath Query Predicate Structures 127

6.2.2 Defining Global Parameters 128

6.3 Semantic Transformations for Predicates 128

6.3.1 Determination of Status for Query Conditions and Their Predicate 128

6.3.2 Predicate Elimination Semantic Transformation 131

6.3.3 Predicate Reduction Semantic Transformation 139

6.3.4 Proposed Functions 144

6.4 Summary 150

Chapter 7 Experimental Design 151

7.1 Experiment Design: A Background 151

7.1.1 Experiment Objectives 152

7.1.2 Experiment Evaluation Objectives 152

7.2 Common Setup for Experiments 153

7.2.1 Implementation Framework & Platform 153

7.2.2 Supporting of Minimal Requirement 154

7.2.3 Choice of Experiment Data and Schema 154

7.2.4 Setup of Operational Hardware, Software and System Modules 156

7.2.5 Computation Procedure and Metrics 161

7.3 Individual Setup for Individual Experiment 162

7.3.1 Experiment 1: Using DBLP Data 162

7.3.2 Experiment 2: Using Benchmark Data 171

7.4 Summary 174

Chapter 8 Experimental Evaluation - Using Real Data Sets 175

8.1 Performance Evaluation Preface 175

8.2 Semantic Path Transformation 176

8.3 Semantic Transformations for Predicates 184

8.4 Semantic Transformation for Axes 193

8.5 Semantic Conflict Detection 202

8.6 Summary 206

Chapter 9 Experimental Evaluation Using Benchmark Data Sets 207

9.1 Performance Evaluation 207

9.1.1 Single Query Condition with Value-based Comparison 208

9.1.2 Twig Pattern Query Conditions with Value-based Comparison 220

9.2 Summary 225

Chapter 10 Conclusions and Future Work 227

10.1 Summary of the Contributions 227

10.2 Limitations of the Work 235

10.3 Future Work 236

10.4 Closing Statement 237

References.. 239

Appendix 1: DBLP XML Schema 251

Appendix 2: Michigan Benchmarking XML Schema 257

List of Figures

Figure 1.1 Example of XML Fragment... 2

Figure 1.2 Outline of Thesis Scope.. 10

Figure 1.3 Inter-relationships between Chapters in the Thesis 13

Figure 2.1 Semantic Query Optimization for Legacy Databases.. 19

Figure 2.2 Summary of Existing Works for XML Query Specified without Predicates 23

Figure 2.3 Summary of Existing Works for XML Query Specified without Predicates 31

Figure 3.1 Overview of Semantic Transformation Methodology 40

Figure 3.2 Location Paths 48

Figure 3.3 Non-target Elements and Location Step, Target Element and Location Step 48

Figure 3.4 An Overview of XML Schema Specification.. 51

Figure 3.5 An Overview of an Identity Constraint using Key and KeyRef 54

Figure 4.1 ComplexType and SimpleType Elements 65

Figure 4.2 Some Common Constraints in XML Schema.. 67

Figure 4.3 Unique Path Derivation using Depth-First Search... 68

Figure 4.4 Inter-relationships of Semantic Path Transformations 71

Figure 4.5 XPathQuery *p* and List of Unique Paths *Q*.. 72

Figure 4.6 XPath Query *p* and list of Unique Paths *Q* 77

Figure 4.7 XPath Query *p* and List of Unique Paths *Q*.. 80

Figure 5.1 Sample of Semantic Transformation with a Bottom-up Approach 89

Figure 5.2 Semantic Transformation for Preceding-sibling Axis Query 93

Figure 5.3 Semantic Transformation for Preceding Axis Query... 100

Figure 5.4 Semantic Transformation for Ancestor or Ancestor-or-self 105

Figure 5.5 Semantic Transformation for Parent Axis Location Step...................................... 109

Figure 5.6 Semantic Transformation for Descendant or Descendant-or-self Axis Query 113

Figure 5.7 Semantic Transformation for *Self* Axis Query.. 118

Figure 5.8 Semantic Transformation for Child Axis Query.. 121

Figure 6.1 Semantic Transformations for XPath Queries Specified with Predicates Framework... 127

Figure 6.2 XPath Query Predicate Holds a Single XPath query Condition with No Value..... 134

Figure 6.3 Condition Status for Predicates in p_1 and p_2 135

Figure 6.4 Schema Fragment Which Exhibits a Keyref Constraint....................................... 136

Figure 6.5 Sample XPath Path and Semantics for Semantic Predicate Reduction.................. 142
Figure 6.6 Condition Status for Predicates.. 142
Figure 7.1 Overview of System Architecture and Task Modules 157
Figure 7.2 DBMP XML Schema and Snapshot of XML Data (after modification) 165
Figure 7.3 Scaled Data Sets for DBLP Data 166
Figure 7.4 Full-path XPath Queries 167
Figure 7.5 Partial-Path XPath Queries 169
Figure 7.6 Micro-benchmark and Snapshot of XML Data 172
Figure 8.1 XPath Queries and Semantic XPath Queries by Semantic Path Transformation .. 177
Figure 8.2 Performance Results Before and After Semantic Path Expansion Applied............ 178
Figure 8.3 Perfomrance Results Before and Afer Semantic Path Contraction Applied........... 180
Figure 8.4 Performance Results Before and After Semantic XPath Compelement Applied ... 182
Figure 8.5 Original and Semantic XPath Queries and Related Information 185
Figure 8.6 Performance Results Before and After Predicate Elimination Applied.................. 186
Figure 8.7 Performance Results Before and After Predicate Elimination Applied.................. 188
Figure 8.8 Performance Results Before and After Predicate Reduction Applied 190
Figure 8.9 Performance Results Before and After Predicate Reduction Applied 191
Figure 8.10 Original and Semantic XPath Queries and Related Information 193
Figure 8.11 Before and After Semantic Transformation for Preceding-/Following-sibling Axis Applied 194
Figure 8.12 Before and After Semantic Transformation for Preceding/Following Axis Applied 196
Figure 8.13 Before and After Semantic Transformation for Ancestor/Ancestor-Or-Self Axis Applied 199
Figure 8.14 Semantic Transformation of XPath Queries with Combination of Axes.............. 202
Figure 8.15 XPath Queries with No Semantic XPath Queries 203
Figure 8.16 Before and After Applied Semantic Conflict Detection of XPath Transformation204
Figure 8.17 Before and After Applied Semantic Conflict Detection Transformation 205
Figure 8.18 Identifiying Semantic Transformations as Optimization Devices 206
Figure 9.1 XPath Queries and Semantic XPath Queries 209
Figure 9.2 Performance Results Before and After Semantic Transformations 209
Figure 9.3 Performance Results Before and After Semantic Transformations 211
Figure 9.4 Performance Results Before and After Semantic Transformations 212
Figure 9.5 XPath Queries and Semantic XPath Queries 214
Figure 9.6 Performance Results Before and After Semantic Transformations 215
Figure 9.7 Performance Results Before and After Semantic Transformations 217

Figure 9.8 Performance Results Before and After Semantic Transformations 219
Figure 9.9 XPath Queries and Semantic XPath Queries 220
Figure 9.10 Performance Results Before and After Semantic Transformations Applied 221
Figure 9.11 Performance Results Before and After Semantic Transformations Applied 223
Figure 9.12 Performance Results Before and After Semantic Transformations Applied 225