



dandelion.com

Sourav S. Bhowmick
Wee Keong Ng

Sanjay K. Maaria

© 2008 AGI-Information Management Consultants
May be used for personal purposes only or by
libraries associated to dandelion.com network.

Web Data Management

A Warehouse Approach

With 106 Illustrations



Springer

Contents

Preface	vii
1 Introduction	1
1.1 Motivation	2
1.1.1 Problems with Web Data	2
1.1.2 Limitations of Search Engines	5
1.1.3 Limitations of Traditional Data Warehouse	7
1.1.4 Warehousing the Web	10
1.2 Architecture and Functionalities	11
1.2.1 Scope of This Book	13
1.3 Research Issues	14
1.4 Contributions of the Book	15
2 A Survey of Web Data Management Systems	17
2.1 Web Query Systems	18
2.1.1 Search Engines	18
2.1.2 Metasearch Engines	20
2.1.3 W3QS	21
2.1.4 WebSQL	27
2.1.5 WebLog	28
2.1.6 NetQL	30
2.1.7 FLORID	32
2.1.8 RAW	35
2.2 Web Information Integration Systems	35
2.2.1 Information Manifold	36
2.2.2 TSIMMIS	37
2.2.3 Ariadne	39
2.2.4 WHIRL	40
2.3 Web Data Restructuring	40
2.3.1 STRUDEL	41
2.3.2 WebOQL	44
2.3.3 ARANEUS	45

2.4	Semistructured Data	47
2.4.1	Lore	47
2.4.2	UnQL	50
2.5	XML Query Languages	50
2.5.1	Lorel	52
2.5.2	XML-QL	56
2.6	Summary	61
3	Node and Link Objects	65
3.1	Introduction	65
3.1.1	Motivation	65
3.1.2	Our Approach - An Overview of WareHouse Object Model (WHOM)	69
3.2	Representing Metadata of Web Documents and Hyperlinks	69
3.2.1	Metadata Associated with HTML and XML Documents	69
3.2.2	Node Metadata Attributes	70
3.2.3	Link Metadata Attributes	70
3.3	Representing Structure and Content of Web Documents	70
3.3.1	Issues for Modeling Structure and Content	72
3.3.2	Node Structural Attributes	74
3.3.3	Location Attributes	79
3.4	Representing Structure and Content of Hyperlinks	80
3.4.1	Issues for Modeling Hyperlinks	81
3.4.2	Link Structural Attributes	82
3.4.3	Reference Identifier	82
3.5	Node and Link Objects	84
3.6	Node and Link Structure Trees	84
3.7	Recent Approaches in Modeling Web Data	87
3.7.1	Semistructured Data Modeling	88
3.7.2	Web Data Modeling	89
3.7.3	XML Data Modeling	89
3.7.4	Open Hypermedia System	90
3.8	Summary	91
4	Predicates on Node and Link Objects	93
4.1	Introduction	94
4.1.1	Features of Predicate	96
4.1.2	Overview of Predicates	97
4.2	Components of Comparison-Free Predicates	100
4.2.1	Attribute Path Expressions	101
4.2.2	Predicate Qualifier	105
4.2.3	Value of a Comparison-Free Predicate	106
4.2.4	Predicate Operators	109
4.3	Comparison Predicates	114
4.3.1	Components of a Comparison Predicate	115
4.3.2	Types of Comparison Predicates	117

4.4	Summary	125
5	Imposing Constraints on Hyperlink Structures	127
5.1	Introduction	127
5.1.1	Overview	129
5.1.2	Difficulties in Modeling Connectivities	129
5.1.3	Features of Connectivities	132
5.2	Components of Connectivities	133
5.2.1	Source and Target Identifiers	134
5.2.2	Link Path Expressions	134
5.3	Types of Connectivities	135
5.3.1	Simple Connectivities	135
5.3.2	Complex Connectivities	135
5.4	Transformation of Complex Connectivities	136
5.4.1	Transformation of Case 1	136
5.4.2	Transformation of Case 2	137
5.4.3	Transformation of Case 3	138
5.4.4	Transformation of Case 4	139
5.4.5	Steps for Transformation	139
5.4.6	Graphical Visualization of a Connectivity	141
5.5	Conformity Conditions	141
5.5.1	Simple Connectivities	141
5.5.2	Complex Connectivities	142
5.6	Summary	142
6	Query Mechanism for the Web	145
6.1	Introduction	145
6.1.1	Motivation	145
6.1.2	Our Approach	149
6.2	Coupling Query	154
6.2.1	The Information Space	154
6.2.2	Components	155
6.2.3	Definition of Coupling Query	166
6.2.4	Types of Coupling Query	169
6.2.5	Valid Canonical Coupling Query	170
6.3	Examples of Coupling Queries	172
6.3.1	Noncanonical Coupling Query	173
6.3.2	Canonical Coupling Query	179
6.4	Valid Canonical Query Generation	181
6.4.1	Outline	181
6.4.2	Phase 1: Coupling Query Reduction	182
6.4.3	Phase 2: Validity Checking	189
6.5	Coupling Query Formulation	190
6.5.1	Definition of Coupling Graph	190
6.5.2	Types of Coupling Graph	191
6.5.3	Limitations of Coupling Graphs	194

6.5.4	Hybrid Graph.....	198
6.6	Coupling Query Results	200
6.7	Computability of Valid Coupling Queries	201
6.7.1	Browser and Browse/Search Coupling Queries.....	202
6.8	Recent Approaches for Querying the Web	203
6.9	Summary.....	205
7	Schemas for Warehouse Data	207
7.1	Preliminaries.....	208
7.1.1	Recent Approaches for Modeling Schema for Web Data	208
7.1.2	Features of Our Web Schema	210
7.1.3	Summary of Our Methodology	212
7.1.4	Importance of Web Schema in a Web Warehouse	213
7.2	Web Schema	214
7.2.1	Definition	214
7.2.2	Types of Web Schema	216
7.2.3	Schema Conformity.....	217
7.2.4	Web Table.....	219
7.3	Generation of Simple Web Schema Set from Coupling Query	221
7.4	Phase 1: Valid Canonical Coupling Query to Schema Transformation	221
7.4.1	Schema from Query Containing Schema-Independent Predicates	222
7.4.2	Schema from Query Containing Schema-Influencing Predicates	223
7.5	Phase 2: Complex Schema Decomposition	225
7.5.1	Motivation	225
7.5.2	Discussion	226
7.5.3	Limitations	227
7.6	Phase 3: Schema Pruning.....	228
7.6.1	Motivation	228
7.6.2	Classifications of Simple Schemas.....	228
7.6.3	Schema Pruning Process	231
7.6.4	Phase 1: Preprocessing Phase	232
7.6.5	Phase 2: Matching Phase	233
7.6.6	Phase 3: Nonoverlapping Partitioning Phase	233
7.7	Algorithm Schema Generator	236
7.7.1	Pruning Ratio	237
7.7.2	Algorithm of GenerateSchemaFromQuery	238
7.7.3	Algorithm for the Construct Partition	240
7.8	Web Schema Generation in Local Operations	246
7.8.1	Schema Generation Phase	246
7.8.2	Schema Pruning Phase.....	248
7.9	Summary.....	249

8	WHOM-Algebra	251
8.1	Types of Manipulation	251
8.2	Global Web Coupling	252
8.2.1	Definition	252
8.2.2	Global Web Coupling Operation	253
8.2.3	Web Tuples Generation Phase	254
8.2.4	Limitations	257
8.3	Web Select	259
8.3.1	Selection Criteria	259
8.3.2	Web Select Operator	260
8.3.3	Simple Web Schema Set	260
8.3.4	Selection Schema	261
8.3.5	Selection Condition Conformity	265
8.3.6	Select Table Generation	265
8.4	Web Project	273
8.4.1	Definition	273
8.4.2	Projection Attributes	273
8.4.3	Algorithm for Web Project	278
8.5	Web Distinct	287
8.6	Web Cartesian Product	288
8.7	Web Join	289
8.7.1	Motivation and Overview	289
8.7.2	Concept of Web Join	291
8.7.3	Join Existence Phase	304
8.7.4	Join Construction Phase When $X_{pj} \neq \emptyset$	315
8.7.5	Joined Partition Pruning	327
8.7.6	Join Construction Phase When $X_j = \emptyset$	330
8.8	Derivatives of Web Join	338
8.8.1	σ -Web Join	338
8.8.2	Outer Web Join	344
8.9	Web Union	350
8.10	Summary	351
9	Web Data Visualization	353
9.1	Web Data Visualization Operators	355
9.1.1	Web Nest	355
9.1.2	Web Unnest	356
9.1.3	Web Coalesce	357
9.1.4	Web Expand	359
9.1.5	Web Pack	360
9.1.6	Web Unpack	362
9.1.7	Web Sort	364
9.2	Summary	365

10	Detecting and Representing Relevant Web Deltas	367
10.1	Introduction	367
10.1.1	Overview	368
10.2	Related Work	369
10.3	Change Detection Problem	371
10.3.1	Problem Definition	371
10.3.2	Types of Changes	372
10.3.3	Representing Changes	372
10.3.4	Decomposition of Change Detection Problem	374
10.4	Generating Delta Web Tables	374
10.4.1	Storage of Web Objects	374
10.4.2	Outline of the Algorithm	375
10.4.3	Algorithm Delta	379
10.5	Conclusions and Future Work	387
11	Knowledge Discovery Using Web Bags	389
11.1	Introduction	389
11.1.1	Motivation	390
11.1.2	Overview	391
11.2	Related Work	392
11.2.1	PageRank	393
11.2.2	Mutual Reinforcement Approach	393
11.2.3	Rafei and Mendelzon's Approach	394
11.2.4	SALSA	395
11.2.5	Approach of Borodin et al.	396
11.3	Concept of Web Bag	397
11.4	Knowledge Discovery Using Web Bags	399
11.4.1	Terminology	399
11.4.2	Visibility of Web Documents and Intersite Connectivity	400
11.4.3	Luminosity of Web Documents	406
11.4.4	Luminous Paths	408
11.4.5	Query Language Design Considerations	413
11.4.6	Query Language for Knowledge Discovery	414
11.5	Conclusions and Future Work	415
12	The Road Ahead	417
12.1	Summary of the Book	417
12.2	Contributions of the Book	420
12.3	Extending Coupling Queries and Global Web Coupling Operation	420
12.4	Optimizing Size of Simple Schema Set	421
12.5	Extension of the Web Algebra	421
12.5.1	Schema Operators	422
12.5.2	Web Correlate	424
12.5.3	Web Ranking Operator	424
12.5.4	Operators for Manipulation at Subpage Level	424
12.6	Maintenance of the Web Warehouse	425

12.7 Retrieving and Manipulating Data from the Hidden Web	425
12.8 Data Mining in the Web Warehouse	426
12.9 Conclusions	427
A Table of Symbols	429
B Regular Expressions in Comparison-Free Predicate Values	431
C Examples of Comparison-Free Predicates	436
D Examples of Comparison Operators	443
E Nodes and Links	445
References	449
Index	459