

Contents

1	Introduction	1
1.1	XML Data Model	1
1.2	Emergence of XML Database	3
1.2.1	Flat File Storage	3
1.2.2	Relational and Object Relational Storage	3
1.2.3	Native Storage of XML Data	4
1.3	XML Query Language and Processing	5
1.4	XML Keyword Search	5
1.5	Book Outline	6
	References	7
2	XML Labeling Scheme	9
2.1	Introducing XML Labeling Scheme	9
2.2	Region Encoding Scheme	10
2.3	Dewey and Extended Dewey Scheme	11
2.3.1	Dewey ID Labeling Scheme	11
2.3.2	Extended Dewey and FST	12
2.4	Dynamic Labeling Scheme	16
2.4.1	Region-Based Dynamic Labeling Scheme	17
2.4.2	Prefix-Based Dynamic Labeling Scheme	18
2.4.3	Prime Labeling Scheme	19
2.4.4	The Encoding Schemes	20
2.5	Summary	31
	References	31
3	XML Data Indexing	33
3.1	Introducing XML Data Indexing	33
3.2	Indexes on XML Tree Structure	34
3.2.1	DataGuides	34
3.2.2	1-Index	39
3.2.3	F&B-Index	43

3.3	Index Based on XML Sequencing	54
3.3.1	PRIX: Indexing and Querying XML Using Prüfer Sequences	54
3.3.2	ViST: A Dynamic Index Method for Querying XML Data by Tree Structures	65
3.3.3	APEX: An Adaptive Path Index for XML Data	75
3.4	Summary	87
	References	88
4	XML Tree Pattern Processing	91
4.1	Introducing XML Tree Pattern Processing	91
4.2	XML Structural Join	92
4.2.1	Tree-Merge Join Algorithms	94
4.2.2	Stack-Tree Join Algorithms	97
4.3	XML Holistic Twig Pattern Processing	103
4.3.1	PathStack	104
4.3.2	TwigStack	108
4.3.3	TwigStackList	112
4.3.4	TJFast	122
4.3.5	Experimental Evaluation	128
4.4	XML Query Processing Based on Various Streaming Schemes	134
4.4.1	Tag+Level Streaming and Prefix-Path Streaming (PPS)....	135
4.4.2	iTwigJoin Algorithm	144
4.5	Summary	155
	References	155
5	Ordered and Generalized XML Tree Pattern Processing	157
5.1	Introducing Ordered and Generalized XML Tree Pattern Processing	157
5.2	XML Ordered Query Processing	158
5.2.1	Data Model and Ordered Twig Pattern	159
5.2.2	XML Ordered Query Processing Algorithm	160
5.2.3	Analysis of OrderedTJ	163
5.2.4	Experimental Evaluation	165
5.3	XML Generalized XML Tree Pattern	167
5.3.1	GTJFast Algorithm	168
5.3.2	Analysis of GTJFast	171
5.3.3	Experiments	173
5.4	Extended XML Tree Pattern	174
5.4.1	Extended Tree Pattern Query	176
5.4.2	Matching Cross	177
5.4.3	Holistic Algorithms	183
5.4.4	Experiments	193
5.5	Summary	200
	References	200

6	Effective XML Keyword Search	203
6.1	Introducing Effective XML Keyword Search	203
6.2	XML Keyword Search Semantics	204
6.2.1	LCA and the Meet Operator	205
6.2.2	MLCA and MLCAS	205
6.2.3	SLCA	206
6.2.4	GDMCT	207
6.2.5	ICA (Interested Common Ancestor) and IRA (Interested Related Ancestors)	209
6.2.6	ELCA (Exclusive Lowest Common Ancestor)	210
6.2.7	VLCA (Valuable Lowest Common Ancestor)	210
6.2.8	MCN	211
6.2.9	Meaningful SLCA	211
6.2.10	LCEA (Lowest Common Entity Ancestor)	213
6.2.11	MLCEA (Meaningful LCEA)	213
6.3	XML Keyword Search Algorithms	213
6.3.1	DIL (Dewey Inverted List) Query Processing Algorithm	213
6.3.2	The Stack Algorithm	216
6.3.3	Basic Multiway-SLCA Algorithm (BMS)	218
6.3.4	Incremental Multiway-SLCA Algorithm (IMS)	220
6.3.5	Indexed Stack Algorithm	221
6.3.6	Stack-Based Query Refinement Algorithm	223
6.4	XML Keyword Search Ranking Strategy	224
6.4.1	TF*IDF Cosine Similarity	225
6.4.2	Data Model	226
6.4.3	XML TF&DF	226
6.4.4	Inferring the Node Type to Search For	227
6.4.5	Inferring the Node Types to Search Via	227
6.4.6	Capturing Keyword Co-occurrence	228
6.4.7	Relevance-Oriented Ranking	228
6.5	Summary	231
	References	231
7	XML Keyword Pattern Refinement	233
7.1	Introducing XML Keyword Pattern Refinement	233
7.2	Related Work	237
7.3	Preliminaries	238
7.3.1	Meaningful SLCA	238
7.3.2	Refinement Operations	240
7.4	Ranking of Refined Queries	242
7.4.1	Similarity Score of a RQ	243
7.4.2	Dependence Score of a RQ	245
7.5	Exploring the Refined Query	247
7.5.1	Problem Formulation	247
7.5.2	Subproblems	247

7.5.3	Notations	247
7.5.4	Initialization	248
7.5.5	Recurrence Function	248
7.5.6	Time Complexity	248
7.6	Content-Aware Query Refinement	249
7.6.1	Partition-Based Algorithm	250
7.6.2	Short-List Eager Algorithm	253
7.7	Experiments	256
7.7.1	Equipment	256
7.7.2	Dataset and Query Set	256
7.7.3	Efficiency	257
7.7.4	Scalability	259
7.7.5	Effectiveness of Query Refinement	261
7.8	Summary	264
	References	264
8	LCRA, XML Keyword Search System, and LotusX, Graphical Query Processing System	267
8.1	Introduction of LCRA and LotusX	267
8.2	LCRA: Search Semantics	268
8.2.1	SLCA and LRA	268
8.2.2	Background and Data Model	269
8.2.3	Search Semantics	270
8.3	LCRA, System Architecture, and Ranking Techniques	272
8.3.1	Tree Model	272
8.3.2	Ranking Techniques	273
8.3.3	System Architecture	274
8.3.4	Overview of Online Demo Features	274
8.4	A Position-Aware XML Graphical Search System with Auto-completion	276
8.4.1	System Features	276
8.4.2	LotusX: Architecture and Algorithms	278
8.5	Summary	282
	References	283
9	Summary and the Road Ahead	285
9.1	Summary of This Book	285
9.2	Future Work	286
9.2.1	Full-Fledged XML Query Engine	286
9.2.2	Directed Graph XML Model	286
9.2.3	Extended Dewey Labeling Scheme for Ordered Query	287
9.2.4	Index Structure Based on TJFast	287
9.2.5	MapReduce-Based XML Twig Pattern Matching	288
	References	288
	Index	289