

Ulf Leser • Felix Naumann

Informationsintegration

Architekturen und Methoden zur Integration
verteilter und heterogener Datenquellen



dpunkt.verlag

Inhaltsverzeichnis

1	Einleitung	3
1.1	Integrierte Informationssysteme	4
1.2	Grundlegende Begriffe	6
1.3	Szenarien der Informationsintegration	9
1.4	Adressaten und Aufbau des Buches	12
2	Repräsentation von Daten	17
2.1	Datenmodelle	18
2.1.1	Das relationale Datenmodell	19
2.1.2	XML-Daten	23
2.1.3	Semistrukturierte Daten, Texte und andere Formate	27
2.1.4	Überführung von Daten zwischen Modellen	32
2.2	Anfragesprachen	34
2.2.1	Relationale Algebra	34
2.2.2	SQL	37
2.2.3	Datalog	38
2.2.4	SQL/XML	41
2.2.5	XML-Anfragesprachen	44
2.3	Weiterführende Literatur	46
3	Verteilung, Autonomie und Heterogenität	49
3.1	Verteilung	51
3.2	Autonomie	54
3.3	Heterogenität	58
3.3.1	Technische Heterogenität	62
3.3.2	Syntaktische Heterogenität	64
3.3.3	Heterogenität auf Datenmodellebene	65
3.3.4	Strukturelle Heterogenität	66
3.3.5	Schematische Heterogenität	70
3.3.6	Semantische Heterogenität	73
3.4	Transparenz	78
3.5	Weiterführende Literatur	80

Inhaltsverzeichnis

4	Architekturen	.83
4.1	Materialisierte und virtuelle Integration	.86
4.2	Verteilte Datenbanksysteme	.91
4.3	Multidatenbanksysteme	.93
4.4	Föderierte Datenbanksysteme	.94
4.5	Mediatorbasierte Informationssysteme	.97
4.6	Peer-Daten-Management-Systeme	.101
4.7	Einordnung und Klassifikation	.104
4.7.1	Eigenschaften integrierter Informationssysteme	.104
4.7.2	Klassifikation integrierter Informationssysteme	.110
4.8	Weiterführende Literatur	.111
 II Techniken der Informationsintegration		113
5	Schema- und Metadatenmanagement	.115
5.1	Schemaintegration	.116
5.1.1	Vorgehensweise	.118
5.1.2	Schemaintegrationsverfahren	.119
5.1.3	Diskussion	.122
5.2	Schema Mapping	.123
5.2.1	Wertkorrespondenzen	.127
5.2.2	Schema Mapping am Beispiel	.129
5.2.3	Mapping-Situationen	.134
5.2.4	Interpretation von Mappings	.137
5.3	Schema Matching	.143
5.3.1	Klassifikation von Schema-Matching-Methoden	.145
5.3.2	Schemabasiertes Schema Matching	.146
5.3.3	Instanzbasiertes Schema Matching	.149
5.3.4	Kombiniertes Schema Matching	.153
5.3.5	Erweiterungen	.155
5.4	Multidatenbanksprachen	.157
5.4.1	Sprachumfang	.158
5.4.2	Beispiele	.159
5.4.3	Implementierung von SchemaSQL	.162
5.5	Eine Algebra des Schemamanagements	.165
5.5.1	Modelle und Mappings	.166
5.5.2	Operatoren	.167
5.5.3	Schemaevolution	.168
5.6	Weiterführende Literatur	.171

6	Anfragebearbeitung in föderierten Systemen.	173
6.1	Grundaufbau der Anfragebearbeitung.	174
6.2	Anfragekorrespondenzen.	184
6.2.1	Syntaktischer Aufbau.	188
6.2.2	Komplexe Korrespondenzen.	189
6.2.3	Korrespondenzen mit nicht relationalen Elementen . . .	194
6.3	Schritte der Anfragebearbeitung.	195
6.3.1	Anfrageplanung.	195
6.3.2	Anfrageübersetzung.	200
6.3.3	Anfrageoptimierung.	201
6.3.4	Anfrageausführung.	205
6.3.5	Ergebnisintegration / Datenfusion.	207
6.4	Anfrageplanung im Detail.	208
6.4.1	Prinzip der Local-as-View-Anfrageplanung.	209
6.4.2	Query Containment.	213
6.4.3	»Answering queries using views«.	224
6.4.4	Global-as-View.	230
6.4.5	Vergleich und Kombination von LaV und GaV.	231
6.4.6	Anfrageplanung in PDMS.	233
6.5	Techniken der Anfrageoptimierung.	234
6.5.1	Optimierungsziele.	234
6.5.2	Ausführungsort von Anfrageprädikaten.	237
6.5.3	Optimale Ausführungsreihenfolge.	241
6.5.4	Semi-Join.	244
6.5.5	Globale Anfrageoptimierung.	245
6.5.6	Weitere Techniken.	247
6.6	Integration beschränkter Quellen.	250
6.6.1	Wrapper.	252
6.6.2	Planung mit Anfragebeschränkungen.	257
6.7	Weiterführende Literatur.	262
7	Semantische Integration.	267
7.1	Ontologien.	269
7.1.1	Eigenschaften von Ontologien.	272
7.1.2	Semantische Netze und Thesauri.	277
7.1.3	Wissensrepräsentationssprachen.	282
7.1.4	Ontologiebasierte Informationsintegration.	288
7.2	Das Semantic Web.	295
7.2.1	Komponenten des Semantic Web.	298
7.2.2	RDF und RDFS.	300
7.2.3	OWL - Ontology Web Language.	311
7.2.4	Informationsintegration im Semantic Web.	312
7.3	Weiterführende Literatur.	313

8	Datenintegration	317
8.1	Datenreinigung	318
	8.1.1 Klassifikation von Datenfehlern.	318
	8.1.2 Entstehung von Datenfehlern.	322
	8.1.3 Auswirkungen von Datenfehlern.	323
	8.1.4 Umgang mit Fehlern.	325
	8.1.5 Data Scrubbing.	326
8.2	Duplikaterkennung.	329
	8.2.1 Ziele der Duplikaterkennung	330
	8.2.2 Ähnlichkeitsmaße.	334
	8.2.3 Partitionierungsstrategien.	340
8.3	Datenfusion.	343
	8.3.1 Konflikte und Konfliktlösung.	344
	8.3.2 Entstehung von Datenkonflikten.	345
	8.3.3 Datenfusion mit Vereinigungsoperatoren.	347
	8.3.4 Join-Operatoren zur Datenfusion.	349
	8.3.5 Gruppierung und Aggregation zur Datenfusion.	352
8.4	Informationsqualität	353
	8.4.1 Qualitätskriterien.	354
	8.4.2 Qualitätsbewertung und Qualitätsmodelle.	356
	8.4.3 Qualitätsbasierte Anfrageplanung.	359
	8.4.4 Vollständigkeit	362
8.5	Weiterführende Literatur.	365
ttl	Systeme	369
9	Data Warehouses	371
9.1	Komponenten eines Data Warehouse.	374
9.2	Multidimensionale Datenmodellierung.	376
9.3	Extraktion - Transformation - Laden (ETL).	382
9.4	Weiterführende Literatur.	387
10	Infrastrukturen für die Informationsintegration	389
10.1	Verteilte Datenbanken, Datenbank-Gateways und SQL/MED	390
10.2	Objektorientierte Middleware.	395
10.3	Enterprise Application Integration.	401
10.4	Web-Services.	404
10.5	Weiterführende Literatur.	407

11	Fallstudien: Integration molekularbiologischer Daten ...	409
11.1	Molekularbiologische Daten.	409
11.2	Attributindexierungssysteme.	414
11.3	Multidatenbanksysteme.	416
11.4	Ontologiebasierte Integration.	418
11.5	Data Warehouses.	420
11.6	Weiterführende Literatur.	423
12	Praktikum: Ein förderierter Webshop für Bücher.	425
12.1	Das Konzept.	425
12.2	Zur Durchführung.	427
12.3	Evaluation.	430
	Literaturverzeichnis.	431
	Index.	455