

Index

- Abox, 159–161, 163–165, 167, 170, 175, 176, 178, 179, 182, 185–190, 192, 194, 195
- absolute path, *see* path
- ACID, 294, 308
- asynchronous, 253, 280, 293, 294, 296, 303, 359, 414
- availability, *see* distributed systems
- BATON, 327, 329–331, 333, 336, 338, 425
- BIGTABLE, 327, 331–334, 337, 338
- bottom-up automaton, *see* tree automaton
- browser, *see* Web browser, 7, 16, 29, 67, 126–128, 130, 232, 236, 238, 241, 249, 250, 280, 401, 410
- Bucket (algorithm), 206–210
- Calabash, *see* XML Calabash
- CAP theorem, 294, 302, 303, 308
- CASSANDRA, 336, 337, 417
- CHORD, 313, 321–325, 329, 330, 333, 336–338
- class
- disjointness, 155
 - hierarchy, 24, 145–147, 152
 - intentional definition, 155, 156
 - intersection, 155, 158
 - union, 155, 158
- clustering, 257, 270, 271, 283, 341, 385
- of graphs, 279, 283
- collaborative filtering, 374, 381, 384–386
- item-based, 374, 381, 384, 385
 - user-based, 374, 381, 385, 386
- concurrency, xiv, 292, 333, 415
- consistency, *see* data consistency
- eventual, 294, 303, 308, 309, 409, 414
 - strong, 294, 295, 302
 - weak, 294
- consistency checking, *see* satisfiability checking
- consistent hashing, *see* hashing, consistent
- constraints
- disjointness, 147, 180, 204
 - domain, 147, 155, 160
 - functionality, 157, 193, 215, 216
 - key, 73, 147, 172, 179, 181, 183, 185, 186, 189, 193, 194
- containment, *see* query, containment
- cosine, 261, 270, 382
- COUCHDB, 341, 385, 400–420
- crawler, *see* Web crawler
- CSS, 249, 278
- damping factor, 275–277, 363
- data
- consistency, 72, 73, 162, 172, 182, 184, 186–188, 194, 195, 217, 218, 220, 222, 229, 292–294, 296, 299, 303, 308, 309, 312, 319, 367, 400, 406, 415
 - reconciliation, 294, 400, 406, 416
 - recovery, 295, 296, 298, 301, 308, 333, 348, 360, 389
 - replication, 265, 292–297, 301–304, 306, 308, 309, 313, 316, 319, 320, 324, 326, 331, 333, 336, 344, 348, 390, 400, 401, 406, 414–417
- data locality, 292, 306, 339, 345
- DBpedia, 238, 239
- deduplication (of Web pages), 251
- deep Web, 196, 280, 281, 283
- delta-coding, 268
- Description Logics
- atomic concept, 160, 162, 164–170
 - axioms, 144, 159–162, 164, 167–169
 - complex concept, 160, 162, 163
 - DL-LITE, 168, 182, 195
 - role, 160–163, 165–170
- Dewey identifier, *see* XML, node
- DHT, 305, 313, 321, 325, 336
- distributed file system, 305, 307, 390
- Distributed Hash Table, *see* DHT

432 Index

- distributed system, 287–290, 292, 294, 295, 299, 301–303, 305, 306, 308, 309, 336, 339, 341, 344, 359, 363, 415
- availability, 294–296, 299, 301, 303, 304, 306, 308, 309, 347
- efficiency, 294, 296, 301–303, 310, 329, 331, 359, 382
- reliability, 296, 299, 304, 306, 308
- scalability, 287, 292, 299, 300, 304, 306, 308, 320, 341, 360, 385, 400, 406
- DL, *see* Description Logics
- DNS, 248, 250, 255, 282
- document vector space, 270
- dynamic type checking, *see* type checking
- DYNAMO, 321, 336, 337
- edit distance, 252, 282
- entailment, 144, 159, 162, 170, 238
- entity, 13, 14, 35, 145, 148, 232, 236–239, 279
- EXIST, xiv, 38, 53, 59, 68, 69, 116–120, 123, 126–129, 417
- Fagin's threshold algorithm, 262, 282, 284
- fail-stop, 301, 304, 324
- failover, 301
- failure, 115, 287, 294, 295, 297–299, 301–303, 305, 306, 308, 309, 319, 320, 324, 328, 333, 335, 340, 341, 345, 347, 348, 389
- fault tolerance, 294, 309, 356, 360
- feed (RSS), 240–242, 280
- First Order Logic, xv, 32, 62, 64–68, 80, 144, 151, 159, 212
- flooding, 304, 305, 336
- FOL, *see* First Order Logic
- GAV, *see* Global-As-Views
- GCI, *see* Description Logics
- GeoNames, 240–242
- GFS, 306–309, 333–335, 346
- Global-As-Views, 198–205, 212, 213, 215, 217, 218, 220–222, 230
- gossiping, 316, 321
- graph mining, 272, 278–280, 282
- grouping, 60, 231, 234, 235, 282, 351, 356
- HADOOP, 305, 309, 337, 341, 349, 363, 385, 387–389, 391, 392, 394–398
- HADOOPDB, 360
- hashing, 251, 253, 282, 313, 314, 337, 344, 357
- consistent, 313, 318–322, 336, 337, 400, 417
- linear, 313–315, 336, 338
- HBASE, 337
- HDFS, 305, 388–391, 396, 397
- HITS, 272, 277, 278, 282, 320
- Holistic Twig Join, 109
- hostname, xv, *see* DNS
- HTML, xv, 3, 7, 8, 19, 21, 28, 29, 83, 127, 197, 231, 232, 239, 243, 249–253, 272, 278, 280–282, 372, 401
- validator, 29, 88, 250
- HTTP, 27, 29, 117, 126, 127, 129, 168, 194, 248, 250, 251, 253, 254, 282, 289, 290, 292, 318, 337, 349, 360, 387, 390, 395, 396, 404–409, 417
- HTTPS, 27, 248, 250
- hyperlink, 249, 251, 272, 280
- HYPERTABLE, 337
- information extraction, 281, 283
- information filtering, 374, 375
- information retrieval, xiii–xvi, 8, 21, 26, 29, 70, 110, 112, 119, 128, 129, 143, 144, 172, 238, 239, 241, 247, 250, 251, 253, 254, 257, 259, 262, 263, 265, 267, 270, 272, 273, 277, 278, 280–283, 285, 304, 310–312, 315, 316, 321, 323–334, 336–339, 346, 364, 366, 367, 370–372, 374, 375, 413, 418
- Internet, xiii, xvi, xvii, 3, 6, 17, 20, 26, 29, 118, 232, 248–250, 278, 279, 282, 289–291, 304, 308, 309, 321, 404
- Inverse Rules algorithm, 212–215
- inverted file, 257–259, 261–265, 267, 270, 272, 276, 282–285, 335, 363, 364, 398
- compression, 258, 259, 267–270, 282, 284, 286, 331, 368
- construction, 40, 41, 65, 79, 155, 185, 195, 203, 208, 212, 214, 225, 240, 265, 281, 282, 335, 363
- maintenance, 265, 270, 284, 313, 316, 318, 321, 326, 331, 333, 334, 336, 337, 400, 410, 419
- inverted index, *see* inverted file
- inverted list, *see* posting list
- IP, 250, 289, 304, 318, 320, 359, 396, 406
- Jaccard coefficient, 252
- JavaScript, 5, 22, 67, 249, 278, 280, 400, 401, 407, 410
- Jena, 169, 194, 236–238
- JSON, 401–404
- keyword query, 254, 256, 257, 261, 262, 270, 278, 281, 285, 364, 374
- latency, 264, 290–292, 301–303, 336, 341, 360
- LAV, *see* Local-As-Views
- linear hashing, *see* hashing, linear
- link farm, 278, 284
- linked data, 239
- load balancing, 304, 306, 320, 325, 326, 328, 356, 401
- local area network, 288
- Local-As-Views, 198, 199, 204–213, 215–218, 221, 222, 229, 230
- locality, *see* data locality
- logging, 295, 296, 335, 348, 352, 360, 361, 389, 397
- LUCENE, 364–373
- MapReduce, 339–349, 351, 356–363, 385, 387, 388, 391–398, 400, 401, 406, 410–413, 417–419, 421, 423, 424, 428
- mashup, 240, 280
- master–master, *see* sharding, 294, 414, 415, 418
- master–slave, *see* sharding, 293, 414

- MathML, 21, 29, 30, 70–72
 mediation, 18, 196–201, 204, 215, 222
 Minicon, 205, 210–212, 215, 216, 218, 221, 229, 230
 Monadic Second-Order logic, 79, 80
 MONETDB, 68, 113
 MONGODB, 341, 417
 MovieLens, 374, 375, 377, 386
 MSO, *see* Monadic Second-Order logic
 MusicXML, 20, 129
- namespace, *see* XML namespace, 14, 15, 35, 49,
 68, 70, 83, 86, 123, 135, 148–150, 233, 237, 239,
 306–308, 404
 prefix, 83, 123, 237
 navigation, 22, 23, 26, 31, 32, 39, 43, 44, 46, 62, 96,
 100, 144, 250, 307, 326, 327, 330, 331
 navigational, *see* navigation
 navigational XPath, 62, 63, 67
 NavXPath, *see* navigational XPath
 NFS, 306, 396
 NoSQL, 294, 295, 309, 336, 341, 417
- OASIS, 87, 92
 OEM, 29, 89–92
 ontology, xvi, 27, 143–145, 147–149, 152, 153, 155,
 159, 161, 168–172, 174–183, 194, 195, 215, 217,
 229, 236, 238, 239
 OPIC, 277, 282
 ORDPATH identifier, *see* XML, node
 overlay network, 289, 304
 OWL, 143, 144, 148, 149, 155–161, 163, 167–171,
 195, 239
- P2P, *see* Peer to Peer
 P2P network, *see* peer-to-peer network
 PageRank, xv, 272–278, 282, 284, 363
 path
 absolute, 64, 248
 expression, 32, 35, 38–44, 53, 63, 64, 66
 relative, 44, 64, 248, 388
 Peer to Peer, 199, 222, 229, 289, 301, 303, 304, 309,
 313, 321, 327, 330, 333, 336
 peer-to-peer network, 288–290, 303, 304, 309, 321,
 323, 327
 structured, 301, 305
 unstructured, 304
 PIGLATIN, 339, 340, 348–359, 361–363, 387, 395,
 398, 399, 424
 pipe, 240, 241
 pipeline, 240–242
 posting list, 257–259, 262–268, 270, 284–286, 364,
 369
 preorder, 24, 31, 95, 101, 103, 104, 131, 132, 134,
 139
 processing instruction, 13, 14, 35
 prologue (XML), 11, 13
- QEXO, 68
 QIZX, 68
 query
 Boolean, 175, 187, 218, 220, 222, 261, 285
 containment, 198–200, 204, 209, 210, 212, 216,
 218, 229
 reformulation, 182, 188–195, 217, 218, 220, 221,
 226–230
 unfolding, 164, 165, 202–204, 212–215, 218, 220,
 230
 query log, 272
- random surfer, *see* PageRank
 ranking, 247, 260, 262, 364, 365, 375
 RDF, 72, 89, 118, 143, 144, 148–156, 169, 171–173,
 176, 178, 188, 194, 195, 237, 239
 semantics, 151
 triple, 148–157, 159, 161, 169, 170, 194, 265, 266
 RDF Schema, *see* RDFS
 RDFa, 239
 RDFS, 144, 148, 149, 151–160, 168–171, 176–183,
 194, 195, 236–239
 Really Simple Syndication, *see* RSS
 recommendation, 10, 28, 67, 68, 91, 144, 172, 241,
 282, 317, 374, 375, 377, 378, 380, 381, 383–386
 reconciliation, *see* data reconciliation
 reformulation, *see* query, reformulation
 regular expression, 14, 54, 74, 75, 78, 81, 85, 231,
 233
 regular language, 74, 76, 78, 87, 91
 relationship, 11, 100, 102, 104, 107, 109, 145,
 147–152, 155, 156, 162, 197, 198, 236, 295
 relative path, *see* path
 Relax NG, 87, 88, 92, 93
 relevance, 210, 260, 261, 272
 reliability, *see* distributed systems
 Remote Procedure Call, *see* RPC, 27, 360
 replication, *see* data replication
 Resource Description Framework, *see* RDF
 REST, 126, 127, 129, 224, 241, 279, 301, 304, 335,
 351, 404, 405, 408, 410, 411, 413
 reverse document order, 46, 53
 robot exclusion, 253, 282
 robot trap, 251, 253
 robots.txt, *see* robot exclusion protocol
 RSS, 19, 240–242, 280
 feed, *see* feed
- satisfiability checking, 162, 163, 168, 169
 saturation algorithm, 238
 SAX, 6, 21–23, 30, 31, 81, 83, 89, 131, 135, 139, 398
 SAXON, 68, 231
 scalability, *see* distributed systems
 Scalable Vector Graphics, *see* SVG
 schematron, 88, 92
 search, *see* information retrieval
 seek time, *see* latency, 264, 291, 292
 semantic heterogeneity, 196
 semantic mapping, 197, 198
 serialization, 4, 5, 7, 9, 21, 28, 68, 88, 100, 401, 416
 Service Oriented Architecture Protocol, *see*
 SOAP
 SGML, 6, 9, 28, 29, 81, 249
 sharding, 301, 315, 319, 346, 417

434 Index

- shared-nothing, 289, 295, 360, 414, 417
- shingle, 252, 253, 282
- Sig.ma, 239
- Simple API for XML, *see* SAX
- sitemap, 251, 282
- SOAP, 6, 27, 28, 359
- Soundex, 256, 282
- spamdexing, 278
- SPARQL, 155, 169, 172, 173, 194, 237, 238
- SQL, 22, 26, 32, 34, 41, 54, 55, 57, 58, 60, 68, 113, 114, 182, 188, 260, 336, 348, 351, 353, 356, 360–362, 374, 375, 377, 378, 383, 400
- STA join, *see* stack-based join
- stack-based join, 104–107
- Standard Generalized Markup Language, *see* SGML
- static type checking, *see* type checking
- STD join, *see* stack-based join
- stemming, 255–257, 282, 285, 365
 - lexical, 256
 - morphological, 255, 356
 - phonetic, 256, 282
 - Porter's, 256
- stop word, 256, 257, 371
- storage balancing, 328
- structural join, 103–109, 112–114
- subsumption, 162, 163, 165–170
- super-peer, 301, 309
- SVG, 19, 20, 29

- tableau method, 163–165, 167, 170
- tableau rules, 164, 165, 167, 170
- taxonomy, 145, 236
- Tbox, 159–165, 167–169, 175, 176, 178, 179, 182–195, 217–220, 222
 - closure, 195
 - NI-closure, 183, 184, 186–188
- TCP, 248, 250, 359
- tf-idf*, 260, 261, 263, 264, 272, 283, 372
- token, 21, 22, 30, 31, 247, 251, 252, 254, 255, 257, 265, 272
- tokenization, 254–257, 259, 368
- top-down automaton, *see* tree automaton
- topology, *see* network, topology
- topology (network), 302, 304, 305, 321, 330, 417
- transaction, 21, 23, 292, 295–299, 302, 303, 308, 341, 353, 359, 360, 409, 415, 416
 - distributed, 296, 298, 299, 360
- transforming XML documents, *see* XML transformation
- tree automaton, 4, 76, 78–80, 82, 87, 91–94
 - bottom-up, 76, 77, 79, 92, 93
 - top-down, 77, 84, 86, 92
- tree pattern, 101, 103, 107–110, 112–115, 131–138, 231
- triple, 172, 173, 176, 179, 194, 237, 238
- TrustRank, 278, 282, 283
- TwigStack join, *see* holistic twig join
- two-phase commit, 297
- type checking, 72, 73, 92, 406
 - dynamic, 72, 73
 - static, 72–74, 91
- unfolding, *see* query, unfolding
- Uniform Resource Identifier, *see* URI
- Uniform Resource Locator, *see* URL
- Uniform Resource Name, *see* URN
- URI, 13, 37, 38, 123, 127, 148–150, 237–239, 321, 404, 405, 407–409
 - cool, 238, 239
- URL, 27, 83, 92, 114, 116, 117, 126, 127, 148, 149, 238–243, 248–251, 253–255, 277, 285, 337, 363, 364, 387, 407, 410
 - absolute, 248
 - fragment, 248
 - query string, 248, 368
 - relative, 248
- URN, 28

- valid document, 15, 21, 72–74, 83, 86
- variable bit encoding, 269
- variable byte encoding, 269
- verification, 35, 73, 74, 82, 172, 205
- VOLDEMORT, 336, 417

- W3C, 10, 28, 29, 32, 34, 42, 67, 68, 72, 80, 83, 84, 87, 88, 91, 144, 171, 172, 238, 240, 241, 249, 250, 282
- Web 2.0, 280
- Web application, 18, 42, 128, 130, 240, 253, 279, 280, 365, 374
- Web browser, 3, 8, 19, 28, 83, 117, 128, 238, 239, 401
- Web client, 16, 250
- Web crawler, 249–251, 253, 254, 267, 276, 277, 279, 280, 283–285
 - ethics, 253
- Web graph, 272, 273, 276–278, 283
- Web robot, *see* Web crawler
- Web server, xiv, xvi, 16, 17, 27, 28, 73, 238, 250, 253, 280, 375, 390, 405
- Web service, xiii, xiv, xvi, 14, 22, 26–28, 84, 123, 126, 148, 339, 359, 360, 396
- Web Service Description Language, *see* WSDL
- Web spider, *see* Web crawler
- well-formed documents, 13, 15, 29, 74, 82, 100, 232, 250, 272
- wget, 238, 239
- word automaton, 75, 76, 78, 92
- workflow, 240, 241, 297, 339, 350, 351, 392
- wrapper, 17, 18, 197, 231, 232, 281
- wrapping, 231, 243
- WSDL, 6, 27, 28, 84

- XHTML, 3, 7, 17, 19, 22, 29, 42, 72, 78, 83, 88, 128, 232–234, 249, 282
- XInclude, 241
- XML fragmentation, 96–98
- XML node
 - attribute node, 35, 47, 52, 70, 131
 - Dewey identifiers, 65, 101–103, 112
 - element node, 35, 44, 47, 48, 54, 55, 70, 131, 148

Cambridge University Press

978-1-107-01243-1 - Web Data Management

Serge Abiteboul, Ioana Manolescu, Philippe Rigaux, Marie-Christine Rousset and Pierre Senellart

Index

[More information](#)

- identifiers, 63, 80, 96, 97, 99–102, 113
 - ORDPATH identifiers, 112, 113
 - root, 35, 37–39, 44, 48, 49, 95, 325, 326, 328, 329, 332, 334
 - sibling, 22, 47, 49, 63, 70, 78, 102, 327, 330
- XML Schema, 14, 35, 53, 61, 68, 72, 75, 80, 82–89, 91, 93, 241
- XML shredding, *see* XML fragmentation
- XML transformation, 17, 42, 231
- XML CALABASH, 241, 242
- XPath, 6, 10, 22, 26, 31, 32, 34, 35, 38–45, 47, 49–55, 61–70, 74, 83, 88, 95, 97–99, 113–115, 119, 120, 122, 123, 127, 134, 231, 233, 235, 241–243, 272, 327
- XPath 1.0, 34, 40, 42, 43, 52–54, 62–64, 66–69, 71, 120, 122
- XPath 2.0, 40, 43, 53, 54, 59, 60, 62, 67, 68, 71, 84, 122, 233, 242
- XProc, 240–243
- XQuery, 6, 10, 18, 22, 26, 32, 34, 35, 37, 38, 40–43, 53–62, 68–71, 73, 74, 83, 84, 113, 118–123, 126, 127, 134, 172, 231, 241, 272
- XSLT, 8, 10, 17, 22, 26, 32, 42, 44, 57, 68, 74, 83, 122, 126, 127, 231–235, 241–243
- template, 41, 44, 232, 233
- XSLT 1.0, 42, 235
- XSLT 2.0, 43, 68, 84, 231, 233, 235
- YAGO, 236–239
- Yahoo! Maps, 240, 242
- YAHOO! PIPES, 240, 241, 280