

Index

- approaches for transaction database storage, 243–245
 - simple transaction storage, 244
 - horizontal storage, 245
 - vertical representation, 245
- Apriori–TID algorithm for generating association mining rules, 282–285
- apriori algorithm, 246–280
- ARFF (attribute relation file format), 32, 33
- association mining, 229–233, 282, 319, 327, 360–362
 - support, 234
 - confidence, 235, 236
 - lift, 237
 - naïve algorithm for finding association rules, 240–242
- CAP theorem, 461, 462
- classification, 65–69, 128–139, 216, 219
 - types, 66
 - working, 67
- classification of unlabeled data, 216
- closed and maximal itemsets, 280–281
 - partitioning clustering, 162–180
 - k-means clustering, 162–195
 - hierarchical clustering algorithms (HCA), 181
 - agglomerative clustering, 182
 - divisive clustering, 195–198
 - density-based clustering, 199
- database segmentation, 24
- data cube, 421–422
- data marts, 398–400
- data mining, 17–22
 - techniques, 23–24
- data mining vs. machine learning, 25–27
- data preprocessing, 55–62
 - its need, 55–57
 - its methods, 58–62
- data warehouse, 394–397
 - defining data warehousing, 395
 - data warehouse architecture, 395
 - benefits of data warehousing, 397
 - comparative study of data warehouse with OLTP and ODS, 401
- data warehouse schema, 405–408
 - star schema, 408, 409, 412
 - snowflake schema, 410, 411
 - fact constellation schema (galaxy schema), 412
 - comparison among star, snowflake and fact constellation schema, 413
- decision tree classifier, 69–110, 128–139, 148–150
- deviation detection, 24
- direct hashing and pruning, (DHP) 285–296
- distance metrics, 157–160
 - euclidean distance, 157–158
 - Manhattan distance, 159
 - Chebyshev distance, 160
- dynamic itemset counting, (DIC)
- fact table, 407
- fingerprinting, 371
- fp growth algorithm, 301–313
 - advantages, 314
- hyperlink induced topic search (HITS) algorithm, 372–376
- information gain, 72, 90
- information theory, 70, 71

468 Index

link analysis, 24	OLAP operations, 432–438
machine learning, 1–14	roll-up, 433
applications, 1	drill down, 435
definition, 5	slice and dice, 437
classification of its algorithms, 5–12	pivot, 438
metrics to assess the quality of classifiers, 119	OLAP servers, 429–432
confusion matrix, 120	relational OLAP, 430
precision, 121, 385	MOLAP, 431
recall, 121, 385	ROLAP vs. MOLAP, 432
F-measure, 122	operational data store, 388–393
naïve Bayes method, 110–117	types of ODS, 390
NoSQL data models, 459, 462	architecture of ODS, 391–392
key-value data model, 454–455	advantages of the ODS, 393
column-family data model, 456	PageRank algorithm, 376, 377, 379
document data model, 457–458	predictive modeling, 24
graph databases, 459–460	reinforcement learning, 14–16
NoSQL	resemblance and containment, 370
Defining NoSQL, 453	R language, 42–54
NoSQL vs. RDBMS, 464	its introduction, 42
OLAP, 416–438	features, 43
defining OLAP, 416	basic operators, 45–47
OLAP applications, 417	search engine, 246, 368, 375–378
OLAP features, 417	precision and recall of search engine, 385
OLAP benefits, 418	split algorithm based on gini index, 90
OLAP strengths, 418	supervised learning, 5, 7, 10, 13, 24
comparison between OLTP and OLAP, 418	unsupervised learning, 10–14, 24, 155
differences between OLAP and data mining, 419	web content mining, 369–370