

# Index

- A-Priori Algorithm, 194, 195, 201  
Accessible page, 169  
Ad-hoc query, 116  
Adjacency matrix, 341  
Adomavicius, G., 320  
Advertising, 16, 97, 186, 261  
Adwords, 270  
Afrati, F.N., 53, 372  
Agglomerative clustering, *see* Hierarchical clustering  
Aggregation, 24, 31, 35  
Agrawal, R., 220  
Alon, N., 144  
Alon-Matias-Szegedy Algorithm, 128  
Amplification, 83  
Analytic query, 50  
AND-construction, 83  
Anderson, C., 320  
Andoni, A., 110  
ANF, *see* Approximate neighborhood function  
ANF Algorithm, 367  
Apache, 22, 54  
Approximate neighborhood function, 367  
Arc, 357  
Archive, 114  
Ask, 174  
Association rule, 187, 189  
Associativity, 25  
Attribute, 29  
Auction, 273  
Austern, M.H., 54  
Authority, 174  
Average, 126  
B-tree, 260  
Babcock, B., 144, 260  
Babu, S., 144  
Backstrom, L., 372  
Bag, 37, 59  
Balance Algorithm, 273  
Balazinska, M., 53  
Band, 70  
Bandwidth, 20  
Basket, *see* Market basket, 184, 186, 187, 216  
Bayes net, 4  
BDMO Algorithm, 251  
Beer and diapers, 188  
Bell, R., 320  
Bellkor's Pragmatic Chaos, 290  
Berkhin, P., 182  
Betweenness, 329  
BFR Algorithm, 234, 237  
BFS, *see* Breadth-first search  
Bi-clique, 335  
Bid, 271, 273, 280, 281  
BigTable, 53  
Bik, A.J.C., 54  
Biomarker, 187  
Bipartite graph, 267, 325, 335, 336  
BIRCH Algorithm, 260  
Birrell, A., 54  
Bitmap, 202  
Block, 11, 19, 162  
Blog, 170  
Bloom filter, 122, 200  
Bloom, B.H., 144  
Bohannon, P., 53  
Boldi, P., 372  
Bonferroni correction, 5  
Bonferroni's principle, 4, 5

- Bookmark, 168  
Boral, H., 373  
Borkar, V., 53  
Bradley, P.S., 260  
Breadth-first search, 329  
Brick-and-mortar retailer, 186, 288, 289  
Brin, S., 182  
Broad matching, 273  
Broder, A.Z., 18, 110, 182  
Bu, Y., 53  
Bucket, 9, 119, 134, 138, 200, 251  
Budget, 272, 279  
Budiu, M., 54  
Burrows, M., 53
- Candidate itemset, 197, 210  
Candidate pair, 70, 201, 203  
Carey, M., 53  
Centroid, 223, 226, 232, 235, 239  
Chabbert, M., 320  
Chandra, T., 53  
Chang, F., 53  
Characteristic matrix, 62  
Charikar, M.S., 110  
Chaudhuri, S., 110  
Checkpoint, 43  
Chen, M.-S., 220  
Child, 329  
Cholera, 3  
Chronicle data model, 143  
Chunk, 21, 210, 238  
CineMatch, 316  
Classifier, 298  
Click stream, 115  
Click-through rate, 265, 271  
Clique, 335  
Cloud computing, 15  
CloudStore, 22  
Cluster computing, 20  
Cluster tree, 246, 247  
Clustera, 38, 52  
Clustering, 3, 16, 221, 305, 321, 326  
Clustroid, 226, 232  
Collaboration network, 324
- Collaborative filtering, 4, 17, 57, 261, 287, 301, 325  
Combiner, 25, 159, 161  
Communication cost, 44, 354  
Community, 321, 332, 334, 351  
Commutativity, 25  
Competitive ratio, 16, 266, 269, 274  
Complete graph, 335, 336  
Compressed set, 238  
Compute node, 20  
Computer game, 295  
Computing cloud, *see* Cloud computing  
Confidence, 187, 189  
Content-based recommendation, 287, 292  
Cooper, B.F., 53  
Coordinates, 222  
Cosine distance, 76, 86, 293, 298  
Counting ones, 132, 251  
Craig's List, 262  
Craswell, N., 285  
Credit, 330  
CURE Algorithm, 242, 246  
Currey, J., 54  
Curse of dimensionality, 224, 248  
Cut, 340  
Cyclic permutation, 68  
Cylinder, 12  
Czajkowski, G., 54
- D. Fotakis, 372  
DAG, *see* Directed acyclic graph  
Darts, 122  
Dasgupta, A., 372  
Data mining, 1  
Data stream, 16, 214, 250, 264  
Data-stream-management system, 114  
Database, 16  
Datar, M., 111, 144, 260  
Datar-Gionis-Indyk-Motwani Algorithm, 133  
Dead end, 149, 152, 153, 175  
Dean, J., 53  
Decaying window, 139, 215

- Decision tree, 298  
Degree, 337, 352  
Degree matrix, 341  
Dehnert, J.C., 54  
del.icio.us, 294  
Deletion, 77  
Deli.cio.us, 325  
Dense matrix, 28  
Density, 231, 233  
Depth-first search, 364  
DeWitt, D.J., 54  
DFS, *see* Distributed file system  
Diameter, 231, 233, 358  
Diapers and beer, 186  
Difference, 30, 33, 38  
Dimension table, 50  
Dimensionality reduction, 308  
Directed acyclic graph, 329  
Directed graph, 357  
Discard set, 238  
Disk, 11, 191, 223, 246  
Disk block, *see* Block  
Display ad, 262, 263  
Distance measure, 74, 221, 327  
Distinct elements, 124, 127  
Distributed file system, 21, 184, 191  
DMOZ, *see* Open directory  
Document, 56, 59, 187, 222, 281, 293, 294  
Document frequency, *see* Inverse document frequency  
Domain, 172  
Dot product, 76  
Dryad, 52  
DryadLINQ, 53  
Dual construction, 326  
Dup-elim task, 40  
 $e$ , 12  
Edit distance, 77, 80  
Eigenvalue, 149, 342  
Eigenvector, 149, 342  
Elapsed communication cost, 46  
Email, 324  
Ensemble, 299  
Entity resolution, 91, 92  
Equijoin, 30  
Erlingsson, I., 54  
Ernst, M., 53  
Ethernet, 20  
Euclidean distance, 74, 89  
Euclidean space, 74, 78, 222, 223, 226, 242  
Exponentially decaying window, *see* Decaying window  
Facebook, 168, 322  
Fact table, 50  
Failure, 20, 26, 39, 40, 42  
Faloutsos, C., 373  
False negative, 70, 80, 209  
False positive, 70, 80, 122, 209  
Family of functions, 81  
Fang, M., 220  
Fayyad, U.M., 260  
Feature, 246, 292–294  
Fetterly, D., 54  
Fikes, A., 53  
File, 20, 21, 191, 209  
Filtering, 121  
Fingerprint, 94  
First-price auction, 273  
Fixedpoint, 84, 174  
Flajolet, P., 144  
Flajolet-Martin Algorithm, 125, 366  
Flow graph, 38  
Fortunato, S., 372  
French, J.C., 260  
Frequent bucket, 200, 202  
Frequent itemset, 4, 184, 194, 197, 336  
Frequent pairs, 194, 195  
Frequent-items table, 196  
Friends, 322  
Friends relation, 49  
Frieze, A.M., 110  
Gaber, M.M., 18  
Ganti, V., 110, 260  
Garcia-Molina, H., 18, 182, 220, 260  
Garofalakis, M., 144

- Gaussian elimination, 150  
Gehrke, J., 144, 260  
Generated subgraph, 335  
Genre, 292, 304, 317  
GFS, *see* Google file system  
Ghemawat, S., 53, 54  
Gibbons, P.B., 144, 373  
Gionis, A., 111, 144  
Girvan, M., 372  
Girvan-Newman Algorithm, 329  
Global minimum, 310  
GN Algorithm, *see* Girvan-Newman Algorithm  
Gobioff, H., 54  
Google, 146, 157, 270  
Google file system, 22  
Google+, 322  
Graph, 42, 321, 322, 351, 357  
Greedy algorithm, 264, 265, 268, 272  
GRGPF Algorithm, 246  
Grouping, 24, 31, 35  
Grouping attribute, 31  
Gruber, R.E., 53  
Guha, S., 260  
Gunda, P.K., 54  
Gyongi, Z., 182
- Hadoop, 25, 54  
Hadoop distributed file system, 22  
Hamming distance, 78, 86  
Harris, M., 317  
Hash function, 9, 60, 65, 70, 119, 122, 125  
Hash key, 9, 280  
Hash Table, 352  
Hash table, 9, 10, 12, 193, 200, 202, 204, 280, 282  
Haveliwala, T.H., 182  
HDFS, *see* Hadoop distributed file system  
Head, 362  
Heavy hitter, 351  
Henzinger, M., 111  
Hierarchical clustering, 223, 225, 243, 306, 327
- HITS, 174  
Hive, 53, 54  
Hopcroft, J.E., 364  
Horn, H., 54  
Howe, B., 53  
Hsieh, W.C., 53  
Hub, 174  
Hyperlink-induced topic search, *see* HITS  
Hyracks, 38
- Identical documents, 99  
IDF, *see* Inverse document frequency  
Image, 115, 293, 294  
IMDB, *see* Internet Movie Database  
Imielinski, T., 220  
Immediate subset, 212  
Immorlica, N., 111  
Important page, 146  
Impression, 262  
In-component, 151  
Inaccessible page, 169  
Index, 10, 352  
Indyk, P., 110, 111, 144  
Initialize clusters, 235  
Insertion, 77  
Interest, 188  
Internet Movie Database, 292, 317  
Intersection, 30, 33, 38  
*Into Thin Air*, 291  
Inverse document frequency, *see* TF.IDF, 8  
Inverted index, 146, 262  
Ioannidis, Y.E., 372  
IP packet, 115  
Isard, M., 54  
Isolated component, 152  
Item, 184, 186, 187, 288, 304, 305  
Item profile, 292, 295  
Itemset, 184, 192, 194
- Jaccard distance, 74, 75, 82, 293  
Jaccard similarity, 56, 64, 74, 169  
Jacobsen, H.-A., 53  
Jagadish, H.V., 144  
Jahrer, M., 320

- Jeh, G., 372  
 Join, *see* Natural join, *see* Multiway join, *see* Star join, 353  
 Join task, 40  
 K-means, 234  
 K-partite graph, 325  
 Kalyanasundaram, B., 286  
 Kamm, D., 317  
 Kang, U., 373  
 Karlin, A., 266  
 Kaushik, R., 110  
 Kautz, W.H., 144  
 Key component, 119  
 Key-value pair, 22–24  
 Keyword, 271, 299  
 Kleinberg, J.M., 182  
 Knuth, D.E., 18  
 Koren, Y., 320  
 Kosmix, 22  
 Krioukov, A., 54  
 Kumar, R., 18, 54, 182  
 Kumar, V., 18  
 Kumar, R., 372  
 Label, 322  
 Lagrangean multipliers, 48  
 Lang, K.J., 372  
 Laplacian matrix, 342  
 LCS, *see* Longest common subsequence  
 Leaf, 330  
 Leiser, N., 54  
 Length, 128, 357  
 Length indexing, 100  
 Leskovec, J., 372  
 Leung, S.-T., 54  
 Lin, S., 111  
 Linden, G., 320  
 Linear equations, 150  
 Link, 29, 146, 160  
 Link matrix of the Web, 175  
 Link spam, 165, 169  
 Livny, M., 260  
 Local minimum, 310  
 Locality, 322  
 Locality-sensitive family, 86  
 Locality-sensitive function, 81  
 Locality-sensitive hashing, 69, 80, 294  
 Logarithm, 12  
 Long tail, 186, 288, 289  
 Longest common subsequence, 77  
 LSH, *see* Locality-sensitive hashing  
 Machine learning, 2, 298  
 Maghoul, F., 18, 182  
 Mahalanobis distance, 241  
 Mahoney, M.W., 372  
 Main memory, 191, 192, 200, 223  
 Malewicz, G., 54  
 Malik, J., 373  
 Manber, U., 111  
 Manhattan distance, 75  
 Manning, C.P., 18  
 Many-many matching, 95  
 Many-many relationship, 184  
 Many-one matching, 95  
 Map task, 22, 23, 25  
 Map worker, 25, 27  
 Map-reduce, 15, 19, 22, 27, 159, 161, 210, 255, 353, 360  
 Market basket, 4, 16, 183, 184, 191  
 Markov process, 149, 152, 348  
 Martin, G.N., 144  
 Master controller, 22, 24, 25  
 Master node, 22  
 Matching, 267  
 Matias, Y., 144  
 Matrix, 27, 35, 36, *see* Transition matrix of the Web, *see* Stochastic matrix, *see* Substochastic matrix, 159, 174, *see* Utility matrix, 308, *see* Adjacency matrix, *see* Degree matrix, *see* Laplacian matrix, *see* Symmetric matrix  
 Matthew effect, 14  
 Maximal itemset, 194  
 Maximal matching, 267  
 Mean, *see* Average  
 Median, 126

- Mehta, A., 286  
Merging clusters, 226, 229, 240, 244, 249, 253  
Merton, P., 18  
Miller, G.L., 373  
Minhashing, 63, 72, 76, 82, 294  
Miniclusler, 238  
Minutiae, 94  
Mirrokni, V.S., 111  
Mirror page, 57  
Mitzenmacher, M., 110  
Moments, 127  
Monotonicity, 194  
Most-common elements, 139  
Motwani, R., 111, 144, 220, 260  
Mulihash Algorithm, 204  
Multiplication, 27, 35, 36, 159, 174  
Multiset, *see* Bag  
Multistage Algorithm, 202  
Multiway join, 46, 354  
Mumick, I.S., 144  
Mutation, 80  
Name node, *see* Master node  
Natural join, 30, 34, 35, 45  
Naughton, J.F., 54  
Navathe, S.B., 220  
Near-neighbor search, *see* locality-sensitive hashing  
Negative border, 212  
Neighbor, 347  
Neighborhood, 357, 366  
Neighborhood profile, 357  
Netflix challenge, 2, 290, 316  
Network, *see* Social network  
Newman, M.E.J., 372  
Newspaper articles, 97, 281, 290  
Non-Euclidean distance, 232, *see also* Cosine distance, *see also* Edit distance, *see also* Hamming distance, *see also* Jacard distance, *see also* Levenshtein distance, *see also* Manhattan distance, *see also* Minkowski distance, *see also* Taxicab distance  
Non-Euclidean space, 246, 248  
Norm, 74, 75  
Normal distribution, 237  
Normalization, 301, 303, 314  
Normalized cut, 341  
NP-complete problem, 335  
O'Callaghan, L., 260  
Off-line algorithm, 264  
Olston, C., 54  
Omiecinski, E., 220  
On-line advertising, *see* Advertising  
On-line algorithm, 16, 264  
On-line retailer, 186, 262, 288, 289  
Open Directory, 166  
OR-construction, 83  
Orthogonal vectors, 224  
Out-component, 151  
Outlier, 223  
Overfitting, 299  
Overture, 271  
Own pages, 170  
Paepcke, A., 111  
Page, L., 145, 182  
PageRank, 3, 16, 27, 29, 40, 145, 147, 159  
Pairs, *see* Frequent pairs  
Palmer, C.R., 373  
Pan, J.-Y., 373  
Parent, 329  
Park, J.S., 220  
Partition, 339  
Pass, 192, 195, 202, 208  
Path, 357  
Paulson, E., 54  
PCY Algorithm, 200, 202, 203  
Pedersen, J., 182  
Perfect matching, 267  
Permutation, 63, 68  
PIG, 53  
Pigeonhole principle, 335  
Piotte, M., 320  
Plagiarism, 57, 187  
Pnutes, 53  
Point, 221, 251  
Point assignment, 223, 234, 328  
Polyzotis, A., 53  
Position indexing, 102, 104

- Positive integer, 138  
 Powell, A.L., 260  
 Power law, 13  
 Predicate, 298  
 Prefix indexing, 101, 102, 104  
 Pregel, 42  
 Principal eigenvector, 149  
 Priority queue, 229  
 Privacy, 264  
 Probe string, 102  
 Profile, *see* Item profile, *see* User profile  
 Projection, 30, 32  
 Pruhs, K.R., 286  
 Puz, N., 53  
 Query, 116, 135, 255  
 R-tree, 260  
 Rack, 20  
 Radius, 231, 233, 357  
 Raghavan, P., 18, 182, 372  
 Rajagopalan, S., 18, 182, 372  
 Ramakrishnan, R., 53, 260  
 Ramsey, W., 285  
 Random hyperplanes, 86, 294  
 Random surfer, 146, 147, 152, 166, 347  
 Randomization, 208  
 Rarest-first order, 281  
 Rastogi, R., 144, 260  
 Rating, 288, 291  
 Reachability, 359  
 Recommendation system, 16, 287  
 Recursion, 40  
 Recursive doubling, 361  
 Reduce task, 22, 24  
 Reduce worker, 25, 27  
 Reed, B., 54  
 Reflexive and transitive closure, 360  
 Reina, C., 260  
 Relation, 29  
 Relational algebra, 29, 30  
 Replication, 21  
 Representation, 247  
 Representative point, 243  
 Representative sample, 119  
 Reservoir sampling, 144  
 Restart, 348  
 Retained set, 238  
 Revenue, 272  
 Ripple-carry adder, 138  
 RMSE, *see* Root-mean-square error  
 Robinson, E., 54  
 Root-mean-square error, 290, 309  
 Rosa, M., 372  
 Rounding data, 303  
 Row, *see* Tuple  
 Rowsum, 246  
 Royalty, J., 54  
 S-curve, 71, 80  
 Saberi, A., 286  
 Sample, 208, 212, 214, 217, 235, 243, 247  
 Sampling, 118, 132  
 Savasere, A., 220  
 SCC, *see* Strongly connected component  
 Schema, 29  
 Schutze, H., 18  
 Score, 93  
 Search ad, 262  
 Search engine, 157, 173  
 Search query, 115, 146, 168, 262, 280  
 Second-price auction, 273  
 Secondary storage, *see* Disk  
 Selection, 30, 32  
 Sensor, 115  
 Set, 62, 100, *see* Itemset  
 Set difference, *see* Difference  
 Shankar, S., 54  
 Shi, J., 373  
 Shim, K., 260  
 Shingle, 59, 72, 97  
 Shivakumar, N., 220  
 Shopping cart, 185  
 Shortest paths, 42  
 Siddharth, J., 111  
 Signature, 62, 65, 72  
 Signature matrix, 65, 70

- Silberschatz, A., 144  
Silberstein, A., 53  
Similarity, 4, 15, 56, 183, 294, 301  
Simrank, 346  
Singleton, R.C., 144  
Singular-value decomposition, 308  
Six degrees of separation, 359  
Sketch, 88  
Sliding window, 116, 132, 139, 251  
Smart transitive closure, 362  
Smith, B., 320  
SNAP, 372  
Social network, 321, 322  
SON Algorithm, 210  
Source, 357  
Space, 74, 221  
Spam, *see* Term spam, *see also* Link spam, 324  
Spam farm, 169, 172  
Spam mass, 172, 173  
Sparse matrix, 28, 63, 64, 159, 160, 288  
Spectral partitioning, 339  
Spider trap, 152, 155, 175  
Splitting clusters, 249  
SQL, 29  
Squares, 356  
Srikant, R., 220  
Srivastava, U., 53, 54  
Standard deviation, 239, 241  
Standing query, 116  
Stanford Network Analysis Platform, *see* SNAP  
Star join, 50  
Stata, R., 18, 182  
Statistical model, 1  
Status, 281  
Steinbach, M., 18  
Stochastic matrix, 149  
Stop clustering, 227, 231, 233  
Stop words, 8, 61, 97, 187, 293  
Stream, *see* Data stream  
String, 100  
Striping, 28, 159, 161  
Strong edge, 324  
Strongly connected component, 151, 364  
Toivonen, H., 220  
Strongly connected graph, 149, 358  
Substochastic matrix, 152  
Suffix length, 104  
Summarization, 3  
Summation, 138  
Superimposed code, *see* Bloom filter, 143  
Supermarket, 185, 208  
Superstep, 43  
Support, 184, 209, 210, 212, 214  
Supporting page, 170  
Suri, S., 373  
Surprise number, 128  
SVD, *see* Singular-value decomposition  
Swami, A., 220  
Symmetric matrix, 343  
Szegedy, M., 144  
Tag, 294, 325  
Tail, 362  
Tail length, 125, 366  
Tan, P.-N., 18  
Target, 357  
Target page, 170  
Tarjan, R.E., 364  
Task, 21  
Taxation, 152, 155, 170, 175  
Taylor expansion, 12  
Taylor, M., 285  
Telephone call, 324  
Teleport set, 166, 167, 172, 348  
Teleportation, 156  
Tendril, 151  
Term, 146  
Term frequency, *see* TF.IDF, 8  
Term spam, 146, 169  
TF, *see* Term frequency  
TF.IDF, 7, 8, 293  
Theobald, M., 111  
Thrashing, 161, 200  
Threshold, 71, 141, 184, 210, 214  
TIA, *see* Total Information Awareness  
Timestamp, 133, 252  
Toivonen's Algorithm, 211

- Tomkins, A., 18, 54, 182  
 Tompkins, A., 372  
 Tong, H., 373  
 Topic-sensitive PageRank, 165, 172  
 Toscher, A., 320  
 Total Information Awareness, 5  
*Touching the Void*, 291  
 Transaction, *see* Basket  
 Transition matrix, 348  
 Transition matrix of the Web, 148, 159, 160, 162  
 Transitive closure, 40, 359  
 Transitive reduction, 364  
 Transpose, 175  
 Transposition, 80  
 Tree, 228, 246, 247, *see* Decision tree  
 Triangle, 351  
 Triangle inequality, 74  
 Triangular matrix, 192, 202  
 Tripartite graph, 325  
 Triples method, 193, 202  
 TrustRank, 172  
 Trustworthy page, 172  
 Tsourakakis, C.E., 373  
 Tube, 152  
 Tuple, 29  
 Tuzhilin, A., 320  
 Twitter, 281  
  
 Ullman, J.D., 18, 53, 54, 220, 260, 372  
 Undirected graph, *see* Graph  
 Union, 30, 33, 37  
 Universal set, 100  
 User, 288, 304, 305  
 User profile, 296  
 Utility matrix, 288, 291, 308  
 UV-decomposition, 308, 317  
  
 Valduriez, P., 373  
 Variable, 128  
 Vassilvitskii, S., 373  
 Vazirani, U., 286  
 Vazirani, V., 286  
 Vector, 27, 74, 78, 149, 159, 174, 175, 222  
  
 Vigna, S., 372  
 Vitter, J., 144  
 Volume (of a set of nodes), 340  
 von Ahn, L., 295, 320  
 von Luxburg, U., 373  
  
 Wallach, D.A., 53  
 Wang, J., 317  
 Wang, W., 111  
 Weak edge, 324  
 Weaver, D., 53  
 Web structure, 151  
 Weiner, J., 18, 182  
 Whizbang Labs, 2  
 Widom, J., 18, 54, 144, 260, 372  
 Wikipedia, 324  
 Window, *see* Sliding window, *see also* Decaying window  
 Windows, 12  
 Word, 187, 222, 293  
 Word count, 23  
 Worker process, 25  
 Workflow, 38, 40, 44  
 Working store, 114  
  
 Xiao, C., 111  
  
 Yahoo, 271, 294  
 Yerneni, R., 53  
 York, J., 320  
 Yu, J.X., 111  
 Yu, P.S., 220  
 Yu, Y., 54  
  
 Zhang, T., 260  
 Zipf's law, 15, *see* Power law  
 Zoeter, O., 285