

Subject Index

1:1 match, 429
1:m match, 429, 435–436

A

abnormal behavior, 511–513
absolute URL, 319
access log, 573
accuracy, 65, 79
active learning, 377
actor, 270
ad click rate, 585
ad relevance, 585
adaBoost, 126
adaptive topical crawler, 341
adjusted cosine similarity, 560
ads, 585
 ad click rate, 585
 ad relevance, 585
 sponsored Ads, 572
 cost-per-impression, 590
 cost-per-click, 585, 590
 cost-per-action, 590
agglomerative clustering, 148–150
 algorithm, 148
 average-link method, 150
 centroid method, 150
 chain effect, 149
 complete-link method, 149–150
 dendrogram, 147
 single-link method, 149
 Ward's method, 150
anchor text, 211, 229, 259
anti-click graph, 581
aperiodic, 281, 283
application server logs, 530

apriori algorithm, 20–26
 algorithm, 20–26
 candidate generation, 22
 join step, 22
 pruning step, 22
 downward closure, 20
 interestingness, 23
 lexicographic order, 20
 rule generation, 24
apriori property, 20
area under the ROC curve, 84–85
ARPANET, 3
aspect, 462
 aspect expression, 462
 aspect extraction, 480, 486–492
 aspect name, 462
 explicit aspect expression, 462
 implicit aspect expression, 462
aspect sentiment classification, 481
aspect-based opinion mining, 464, 480
aspect-based opinion summary,
 467–469
association rule, 6, 17–41, 98
 confidence, 18
 minimum confidence, 19
 minimum support, 19
 support, 18,
 support count, 18
 transaction, 18
association rules based
 recommendations, 559, 561–564
associative classification, 93–99
asymmetric attribute, 154
AUC, *see* area under the ROC curve
authority, 5, 288–294
authority ranking, 288, 294

automatic data extraction, 363–364
 automatic wrapper generation,
 381–383
 average precision, 225
 average-link method, 150

B

back-crawl, 350
 backward rule, 377
 bag of words, 215
 bagging, 126
 base type, 477
 base URL, 319
 batch gradient descent, 567, 570–571
 Bayesian Sets, 187
 beam search, 90
 behavioral pattern, 11
 best-first crawler, 315
 best-N-first, 330, 340
 betweenness centrality, 272–273, 301
 Biased-SVM, 197–199
 biases, 569
 bibliographic coupling, 275–276, 277,
 292–293
 binary attribute, 152, 157
 binary split, 75
 bipartite core community, 296–298
 bipartite sub-graph, 288
 bitwise, 237
 bogus opinions, 506
 Boolean query, 213, 216
 boosting, 126–127
 bootstrap replicate, 126
 Borda ranking, 255–256
 breadth-first crawler, 313–314
 breakeven point, 227
 browser, 1
 but-clause, 482

C

candidate itemset, 21
 cannot-link, 166
 canonicalization, 319
 CAR, *see* class association rule

CAR-Apriori, 38–39
 case of letter, 229
 categorical, 26
 CBA, 94
 center, 136, 297
 center star method, 390–391
 central actor, 271
 centrality, 270
 betweenness, 272–273, 301
 closeness, 272
 degree, 271
 centroid, 136, 545
 Chebychev distance, 152
 citation analysis, 275–277
 co-citation, 276
 bibliographic coupling, 277
 class association rule, 36–41, 57,
 94–99
 algorithm, 38–40
 class label, 36
 condset, 38
 condsupCount, 38
 confidence, 37, 38
 confident, 38
 frequent ruleitems, 38
 multiple class supports, 41
 multiple item supports, 41
 ruleitems, 38
 rulesupCount, 38
 support, 37, 38
 class prior probability, 105
 class sequential rule, 55–56, 94
 classification, 63–128
 classification based on association,
 94–98
 association rule, 6, 17–41, 98
 CAR, *see* class association rule
 CBA, 94–98
 class association rule, 94–98
 CMAR, 98
 classifier building, 97
 rules as feature, 98
 strongest rule, 97
 classification model, 64
 classifier, 64
 classifier evaluation, 79–87

- click graph, 578
- clicked documents, 577
- clickstream, 17, 527
- clickthrough data, 539, 578, 583
- clickthrough rate, 585, 590–591
- client-server, 1
- cloaking, 261, 354
- close tag, 368
- closeness centrality, 272
- cluster evaluation, 159–162
 - confusion matrix, 160
 - entropy, 160
 - ground truth, 160
 - indirect evaluation, 162
 - inter-cluster separation, 162
 - intra-cluster cohesion, 162
 - purity, 160
 - user inspection, 160
- cluster of arbitrary shape, 146
- cluster, representation of, 144–146
- clustering, 6, 8, 133–165
- CMAR, 98
- co-citation, 275, 276, 292–293
- co-citation matrix, 276
- co-clustering, 166
- collaborative filtering, 540, 555–571
 - k*-nearest neighbor, 559–561
 - association rule based
 - recommendation, 561–564
 - matrix factorization, 565–571
 - singular value decomposition, 565–571
- collapse, 407
- CombANZ, 255
- combating spam, 262–263
- CombMAX, 254
- CombMIN, 254
- CombMNZ, 255
- CombSUM, 254
- community discovery, 270, 292, 294–303
 - bipartite core, 296–298
 - manifestation, 296
 - email, 301–302
 - maximum flow, 298–301
 - overlapping, 302–303
 - sub-community, 295
 - sub-theme, 295
 - super-community, 295
 - theme, 295
- comparative, 493
 - gradable comparison, 494
 - equative, 494
 - non-equal gradable, 494
 - non-gradable comparison, 494
 - superlative, 494
- comparative opinions, 463
- comparative relations, 494
- comparative type, 477
- complementarity condition, 114, 115, 119
- complete-link method, 149–150
- composite attribute, 433
- composite domain, 433
- computational advertising, 589–593
 - content match-based advertising, 590
 - cost-per-impression, 590
 - cost-per-click, 585, 590
 - cost-per-action, 590
 - online advertising, 589
 - pay per impression, 590
 - sponsored search advertising, 585, 590
 - targeted advertising, 590
- concurrency, 322
- conditional independence assumption, 100–101, 177
- Condorcet ranking, 255, 256
- confidence, 18, 54, 69
- conflict resolution, 405
- confusion matrix, 81, 152, 161
- connectionist reinforcement learning, 345
- consecutive query similarity graph, 579
- constrained optimization problem, 189
- constraint based match, 430–431
- content data, 532
- content hiding, 261
- content match-based advertising, 590
- content spamming, 258–259

content-based recommendations, 557
 content-enhanced transaction matrix, 542
 context-focused crawler, 329–330
 contextual query models, 588
 contiguous sequence, 551
 contiguous sequential pattern, 551
 conversion rate, 591
 co-occurrence, 17
 cookies, 530, 535, 575, 578
 coreference resolution, 501–502
 corpus-based approach, 478–479
 correlation, 337
 cosine similarity, 154, 218
 cosine-rocchio, 192–194
 cost-per-impression, 590
 cost-per-click, 585, 590
 cost-per-action, 590
 co-testing, 377
 co-training, 176–178
 covariate shift, 187
 coverage, 99, 326
 crawl history, 314
 crawler, 10, 311–358

- concurrency, 322–323
- crawl history, 314
- evaluation, 348–353
- fetching, 315
- focused crawler, 311, 327–330
- freshness, 326
- frontier, 312
- live crawling, 356
- page repository, 321–322
- parsing, 316–318
- preferential crawler, 311, 314–315
- robot, 311
- robots.txt, 353
- scalability, 324–326
- spider, 311
- spider trap, 320
- topic crawler, 298, 330
- universal crawler, 311, 314, 323–327

 crawler ethics, 353–355
 crawler etiquette, 353
 crawler evaluation, 348–353

cross-validation, 80
 CSR, *see* class sequential rules
 customer conversion ratios, 539

D

damping factor, 285
 dangling page, 281
 data fusion, 533–534
 data integration, 418
 data mining process, 6
 data pre-processing, 66
 data record, 363, 366, 368, 404–405
 data region, 163, 364, 398, 400
 data sequences, 42
 data standardization, 155–157

- asymmetric binary attribute, 157
- interval-scaled attribute, 155
- mean absolute deviation, 156
- nominal attribute, 152, 157
- ordinal attribute, 157
- range, 155
- ratio-scaled attribute, 157
- symmetric binary attribute, 157
- z-score, 155, 156

 data value match, 418
 decision boundary, 110
 decision list, 87
 decision surface, 110
 decision tree, 67–79

- algorithm, 70
- binary split, 75
- C4.5, 67
- continuous attribute, 75, 76
- decision nodes, 67
- divide-and-conquer, 70
- entropy, 72
- impurity function, 71–75
- information gain, 72, 73
- information gain ratio, 72, 74–75
- leaf nodes, 67
- missing value, 78–79
- overfitting, 76–77
- post-pruning, 77
- pre-pruning, 77
- rule pruning, 78

- skewed class distribution, 79
- split info, 75
- stopping criteria, 70
- tree pruning, 76
- decreasing comparative, 497
- deep Web, 319, 425, 438
- default class, 78, 88, 97
- degree centrality, 271
- degree prestige, 274
- DeLa, 420
- demographic data, 530
- dendrogram, 147
- denial of service, 321
- dense region, 163
- dependency, 471
- dependency grammar, 490
- DEPTA, 420
- description match, 430
- detail page, 364, 413
- Dice function, 442
- dictionary based approach, 477
- direct dependency, 490
- direct opinions, 464
- directory cloning, 260
- discretization, 102
- discriminative model, 201
- disk version of k -means, 139
- distance function, 124, 135, 151–155
 - Chebychev distance, 152
 - cosine similarity, 218
 - Euclidean distance, 151, 152
 - Jaccard coefficient, 154
 - Manhattan distance, 152
 - Minkowski distance, 151
 - simple matching distance, 153
 - squared Euclidean distance, 152
 - weighted Euclidean distance, 152
- distributed hypertext system, 2
- distributional similarity, 187
- divide-and-conquer, 93
- divisive clustering, 148
- document collection, 213
- document index, 215
- Document Object Model, 294, 384, 396

- document-level sentiment
 - classification, 469
- DOM tree, 294, 384, 396–397
- DOM, *see* Document Object Model
- domain matching, 426, 431–434, 441–442
- domain similarity, 441
- double propagation, 490
- downward closure property, 20, 29
- dual, 116
- dual variable, 116
- duplicate detection, 231–232
- duplicate page, 231
- duplicate removal, 254

E

- eager learning, 124
- e-commerce data mart, 539
- edit distance, 384–385
- eigensystem, 275, 279
- eigenvalue, 279, 281
- eigenvector, 279
- Elias Delta coding, 237, 239
- Elias Gamma coding, 237, 238
- EM algorithm, *see* Expectation–Maximization Algorithm
- email community, 301
- emotion, 466
- empty cluster, 138
- empty region, 163
- end rule, 371–372
- ensemble of classifiers, 125
 - bagging, 126
 - boosting, 126–127
 - bootstrap replicate, 126
- entity, 461, 464
 - entity expression, 462
 - entity extraction, 499
 - entity name, 462
- entropy, 72, 160
- episode identification, 537
- equative comparisons, 494
- error rate, 79
- ethical issue, 311

Euclidean distance, 152
 exact match, 216
 exclusion operator, 216
 Expectation–Maximization (EM),
 173, 194, 202–204, 548
 explicit aspect expression, 462
 extraction problem, 382

F

fake opinions, 506
 fake reviews, 506
 false negative, 81
 false positive, 81
 false positive rate, 83
 feature, 462
 feature space, 120
 feature-based opinion mining, 464
 feature-based opinion summary, 467
 flat data record, 368
 flat relation, 367
 flat set type, 367–368
 flat tuple type, 367
 focused crawler, 327–330
 classification, 327
 context graph, 329
 context-focused crawler, 329
 distiller, 329
 Open Directory Project, 327
 forward rule, 377
 frequent itemset, 20, 550
 frequent itemset graph, 563
 frequent sequence, 42
 freshness, 326
 frontier, 312
 F-score, 82, 227
 full document query, 214
 fundamental assumption of machine
 learning, 66

G

gap, 237
 Gaussian fields, 181–182
 Gaussian RBF kernel, 123

generalized node, 400–402
 generative model, 104
 global query interface, 425, 439,
 450–453
 grouping constraint, 451
 instance appropriateness, 453–454
 lexical appropriateness, 452–453
 structure appropriateness, 450–452
 Golomb coding, 240
 Golomb–Rice coding, 241
 gradable comparisons, 494
 grammar induction, 409
 group spammers, 507–508
 group spamming detection, 513
 GSP algorithm, 43, 44

H

harvest rate, 349
 head-item problem, 35
 hidden Web, 319
 hiding technique, 261, 507–508
 Hierarchical clustering, 135, 147
 hit, 473
 HITS, 288–294
 authority, 5, 288–294
 community, 292
 hub, 288–291
 Hypertext Induced Topic Search,
 288
 relation with co-citation and bib-
 liographic coupling, 292–293
 holdout set, 79
 hole, 163
 homonym, 431
 HTML, 2
 HTTP, 2
 hub, 277, 288, 294
 hub ranking, 288–289
 hybrid methods, 557
 hyperlink, 2, 6, 211, 278
 hypermedia, 2
 hypernym, 430
 hypertext, 2
 Hypertext Induced Topic Search, 288

I

- IEPAD, 406
- implicit aspect, 501
- implicit aspect expression, 462
- implicit feedback, 222
- impressions, 539, 590
- impurity function, 71
- inclusion operator, 216
- increasing comparative, 497
- index compression, 237–242
 - Elias Delta coding, 237, 239
 - Elias Gamma coding, 237–238
 - fixed prefix code, 240
 - Golomb coding, 237, 240
 - Golomb-Rice coding, 241
 - integer compression, 237
 - unary coding, 238
 - variable-bit, 237
 - variable-byte, 237, 242
- index construction, 235–236
- indexer, 213, 214
- indexing, 250
- indirect opinions, 464
- individual spammer, 507–508
- inductive learning, 63
- influence domain, 274
- information gain, 72–74, 92
- information gain ratio, 72, 74
- information integration, 425
 - conflict, 445
 - domain similarity, 442
 - global interface, 439
 - grouping relationship, 444
 - h-measure, 446
 - homonym, 431
 - intersect-and-union, 453
 - matcher, 434
 - matching group, 444, 445
 - mutual information measure, 449
 - name as value, 433
 - negatively correlated, 444
 - occurrence matrix, 448
 - positively correlated, 444
 - synonym group, 444
 - transitivity, 444
- information need, 572
- information retrieval, 9, 237–264
- information retrieval evaluation, 223–226
 - average precision, 225
 - breakeven point, 227
 - F-score, 227
 - precision, 224
 - precision-recall curve, 225
 - recall, 224
 - rank precision, 227
- information retrieval query, 213–214
 - Boolean query, 213
 - full document query, 214
 - keyword query, 213
 - multi-word query, 252
 - natural language question, 214
 - phrase query, 213–214
 - proximity query, 214
 - single word query, 252
- information theory, 72
- informative example, 377, 379
- InfoSpiders, 330, 344
- infrequent class, 86
- in-link, 251, 271
- in-link spamming, 260
- input space, 120
- input vector, 109
- instance-based wrapper learning, 378
- instance-level matching, 426, 431
- integer compression, 237
- inter-cluster separation, 162
- interestingness, 23, 57
- Internet, 1, 3
- inter-site schema matching, 449
- interval-scaled attribute, 155
- intra-cluster cohesion, 162
- intra-site schema matching, 449
- inverse document frequency, 217
- inverted index, 215, 232
- inverted list, 233
- IR score, 252
- irreducible, 281, 283, 284
- irregular comparatives, 493
- irregular superlatives, 493
- is-a type, 429, 435

item prediction, 557
 item-based collaborative filtering, 560
 item-based recommendations, 559
 itemset, 18
 iterative SVM, 195

J

Jaccard coefficient, 154, 232, 300
 Jaccard distance, 154

K

KDD, *see* knowledge discovery in database
 kernel function, 111, 120–123
 kernel trick, 123
 keyword query, 211, 213
 keywords, 9, 237
k-means clustering, 136–139

- center, 136
- centroid, 136
- data space, 136
- Euclidean space, 137
- seed, 136
- mean, 137
- outlier, 140

k-modes, 140
k-nearest neighbor, 124
k-nearest neighbor recommendations, 559
k-neighborhood, 580
 knowledge discovery in database, 6
k-sequence, 42
 Kuhn–Tucker conditions, 114, 118

L

label sequential rule, 54
 landmark, 371
 language model, 219, 223
 language pattern, 17
 Laplace smoothing, 103, 107, 220
 latent factor models, 565
 latent semantic indexing, 243–249

- k*-concept space, 245

- left singular vector, 244
- query and retrieval, 236–237
- right singular vector, 244
- singular value decomposition, 243, 246–247

latent variables, 565
 lazy learning, 124
 learning algorithm, 65
 learning from labeled and unlabeled examples, 171–184, 222
 learning from positive and unlabeled examples, 171, 184–202, 222
 learning rate, 567
 learning to rank, 587
 learn-one-rule, 90–93
 least commitment approach, 392
 leave-one-out cross-validation, 80
 level-wise search, 20
 lexicographic order, 20, 42
 lexicon-based approach, 481
 Lidstone smoothing, 103, 107, 220
 lifetime values, 539
 lift chart, 86
 lift curve, 86
 likely positive set, 190
 linear learning system, 109
 linear SVM: non-separable case, 117–120
 linear SVM: separable case, 111–116
 linguistic pattern, 17
 linguistic similarity, 442
 link analysis, 269
 link canonicalization, 318
 link extraction, 318
 link spamming, 259–260
 link topology, 333
 linkage locality, 336
 link-cluster conjecture, 333
 link-content conjecture, 333
 list iteration rule, 370
 list page, 360, 413
 live crawling, 356
 longest common subsequence, 406
 longest repeating subsequences, 552
 LSI query and retrieval, 246
 LSI, *see* latent semantic indexing

LSR, *see* label sequential rule
 LU learning, 171–184, 558
 co-training, 176–178
 combinatorial Laplacian, 182
 constrained optimization, 189
 EM-based algorithm, 173–174,
 202–204
 evaluation, 184
 Gaussian fields, 182
 mincut, 181
 self-training, 178–179
 spectral graph transducer, 182–183
 theoretical foundation, 187–189
 transductive SVM, 179–180
 transduction, 179
 weighting the unlabeled data, 175

M

m:n, 429
 main content block, 5
 Manhattan distance, 151, 152
 MAP, *see* maximum a posteriori
 margin, 111, 112
 margin hyperplane, 112
 market basket analysis, 17
 Markov chain, 279
 Markov models, 551
 match cardinality, 429
 matcher, 434
 matching group, 444, 445
 matrix factorization for
 recommendations, 559, 565–571
 MaxDelta, 438
 maximal margin hyperplane, 111
 maximum a posteriori, 100
 maximum flow community, 298–301
 maximum likelihood estimation, 202
 maximum matching, 387
 maximum support difference, 30
 MDR, 402
 mean absolute deviation, 156
 Mercer's theorem, 123
 meta-search, 253–257
 Borda ranking, 255–256
 CombANZ, 255
 combine similarity scores, 254
 CombMAX, 254
 CombMIN, 254
 CombMNZ, 255
 CombSUM, 254
 Condorcet ranking, 255–256
 duplicate removal, 253
 fuse, 253
 merge, 253
 reciprocal ranking, 256
 minconf, 19
 mincut, 181
 minimum class support, 41, 96
 minimum confidence, 19
 minimum item support, 28, 29, 41
 minimum support, 19, 29, 42, 45
 mining comparative opinions,
 493–498
 Minkowski distance, 151
 minsup, 19
 mirror site, 231
 mirroring, 231
 MIS, *see* minimum item support
 Missing references, 538
 missing value, 78–79, 103
 mixture component, 104
 mixture model, 104
 mixture models, 547
 mixture of Markov models, 548
 mixture probability, 104
 mixture weight, 104
 Mosaic, 3
 MS-Apriori, 30
 MS-GSP, 46
 MS-PS, 52
 multinomial distribution, 106, 108
 multinomial trial, 106
 multiple alignment, 390–396
 center star method, 390–391
 partial tree alignment, 391–396
 multiple minimum class supports, 41,
 96
 multiple minimum item supports, 41
 multiple minimum supports, 26–36,
 45–49, 52, 54
 algorithm, 32

- downward closure, 29
- extended model, 28–30
- head-item, 35
- join step, 33
- minimum item support, 29
- prune step, 33
- rare item, 27, 28
- rule generation, 35–36
- multiple minimum supports, 564
- multiple random sampling, 80
- multivariate Bernoulli distribution, 108
- multi-word query, 252
- must-link, 166
- mutual information measure, 449
- mutual reinforcement, 288

N

- naïve Bayesian classification, 99–103
 - assumption, 100
 - Laplace’s law of succession, 103
 - Lidstone’s law of succession, 103
 - maximum a posteriori (MAP), 100
 - missing value, 103
 - numeric attribute, 102
 - posterior probability, 100
 - prior probability, 100
 - zero count, 102–103
- naïve Bayesian text classification, 103–109
 - assumption, 106
 - generative model, 104
 - hidden parameter, 104
 - mixture model, 104–105
 - mixture component, 104
 - mixture probability, 104
 - mixture weight, 104
 - multinomial distribution, 106–107
 - multivariate Bernoulli distribution, 108
- naïve best-first, 339
- name match, 429–430
- named entity community, 302
- navigational pattern, 550
- nearest neighbor learning, 180

- negative potential items, 484
- negatively correlated, 444
- nested data record, 407
- nested relation, 366
- NET, 407–408
- Netflix Prize contest, 565
- Netscape, 3
- neutral, 411
- n-fold cross-validation, 80
- n-gram, 232
- nominal attribute, 152, 154, 156
- non-contextual query models, 588
- non-gradable comparisons, 494
- nonlinear SVM, 120
- normal vector, 111
- normalized edit distance, 386
- normalized term frequency, 217
- normalized tree match, 389

O

- occurrence type, 251–252
- ODP, *see* Open Directory Project
- Okapi, 218
- online advertising, 589
- Open Directory Project, 327, 583
- open tag, 368
- opinion context, 479
- opinion definition, 463
- opinion holder, 462
- opinion holder extraction, 498
- opinion idioms, 471
- opinion lexicon, 477, 481
- opinion mining, 459–514
- opinion orientations, 463
- opinion phrases, 470
- opinion retrieval, 503–506
- opinion search, 503–506
- opinion shifters, 482
- opinion sources, 463
- opinion spam, 506–514
- opinion spam types, 506
- opinion spam detection, 509–514
- opinion target, 461
- opinion words, 470, 477
- opinionated text, 459

opinion-bearing words, 470, 477
 optimal cover, 346
 ordinal attribute, 157
 orthogonal iteration, 292
 outlier, 140
 out-link, 271
 out-link spamming, 259
 overfitting, 76–78, 569
 overlapping community, 302–303

P

packet sniffer, 538, 573
 page content, 7
 page repository, 321–322
 PageRank, 10, 277–288
 aperiodic, 281–284
 damping factor, 285
 irreducible, 281–284
 Markov chain, 279–281
 power iteration, 279, 285
 principal eigenvector, 279, 281
 random surfer, 280
 stationary probability distribution, 281
 stochastic transition matrix, 281
 strongly connected, 283
 pageview, 531, 540
 pageview identification, 529, 534
 pageview-feature matrix, 542
 pageview-weight, 546
 partial tree alignment, 391–396, 405
 partially supervised learning, 171
 partitional clustering, 135–136
 part-of-speech, 471
 path completion, 538–539
 pay per impression, 590
 Pearson's correlation coefficient, 559
 Penn Treebank POS tags, 472
 personalized Web content, 545
 phrase query, 213–214
 pivoted normalization weighting, 219
 PLSA, *see* Probabilistic Latent Semantic Analysis
 PMI, *see* pointwise mutual information
 pointwise mutual information, 473
 polar words, 477
 polarity, 463
 polynomial kernel, 122–123
 polysemous query, 572
 POS, *see* part-of-speech
 positive potential items, 484
 positively correlated, 444
 post-pruning, 77
 power iteration, 285
 precision, 81–83, 224, 311
 precision and recall breakeven point, 82–83
 precision-recall curve, 225
 predictive model, 64
 preferential crawler, 311, 314–315
 preferred entity, 496–498
 PrefixSpan algorithm, 50–51
 pre-processing, 528–540
 pre-pruning, 77
 prestige, 270, 273–275
 degree prestige, 274
 proximity prestige, 274
 rank prestige, 275, 278
 primal, 115
 primal variable, 115, 119
 principal eigenvector, 279, 281
 Probabilistic Latent Semantic Analysis, 548
 profile, 11
 prominence, 275
 proper subsequence, 54
 proximity prestige, 274
 proximity query, 214
 proxy logger, 573
 pseudo-relevance feedback, 223
 PU learning, 171, 185, 222–223, 558
 biased-SVM, 197–199
 classifier selection, 196–197, 199
 constrained optimization, 189
 direct approach, 190
 EM algorithm, 194–195
 evaluation, 201
 IDNF, 194
 iterative SVM, 195–196
 Probability estimation, 199–201

- Rocchio classification, 220
- S-EM, 191–192, 205
- Spy technique, 191–192
- theoretical foundation, 187–190
- two-step approach, 190
- reliable negative, 190
- purity, 160

Q

- QLM, *see* query log mining
- quality page, 251
- query, 213–215
 - Boolean query, 213
 - full document query, 214
 - keyword query, 213
 - multi-word query, 252
 - natural language question, 214
 - phrase query, 213–214
 - proximity query, 214
 - single word query, 252
- query classification, 586
- query expansion, 583
- query log data models, 577–581
- query log data preparation, 575–577
- query log features, 582–583
- query log mining, 571–589
- query logs, 571
- query operation, 214
- query recommendation, 583
- query refinement, 584
- query suggestion, 583
- query-graph features, 582

R

- random surfer, 280
- rank precision, 227
- rank prestige, 273, 275
- ranking SVM, 223
- rare classes, 86
- rare item problem, 27, 28
- rating prediction, 556–557
- ratio-scaled attribute, 156
- recall, 81, 189, 224, 349
- receiver operating characteristics, 83–85

- recency search, 286–288
- reciprocal ranking, 256
- recommendation engine, 11, 528
- recommender systems, 555–571
- redirection, 261–262
- redundant rule, 39
- regular expression, 382, 409–412, 415–416,
- regular opinions, 463
- regularization, 569
- regularization constant, 569
- reinforcement learning, 343
- re-labeling, 378
- related queries, 584
- relative URL, 319
- re-learning, 378
- relevance feedback, 214, 220–223
- reliable negative document, 190
- replication, 231
- reputation score, 252
- reuse of previous match results, 436–437
- right singular vector, 244
- RMSE, *see* root mean square error, 565
- RoadRunner, 414–415
- robot, 311
- robot exclusion protocol, 353
- robot visits, 576–577
- robots.txt, 353
- ROC curve, 83–85
- Rocchio classification, 221–222
- Rocchio method, 558
- Rocchio relevance feedback, 221
- root mean square error, 565
- rule induction, 87–93
 - decision list, 87
 - default class, 78
 - ordered class, 88–89
 - ordered rule, 88
 - rule pruning, 93
 - separate-and-conquer, 93
 - understandability, 93
- rule learning, 87
- rule pruning, 78, 93, 96
- rule understandability, 93

ruleitem, 38
 rules of opinions, 483–486

S

sample selection bias, 187
 scale-up method, 151
 schema matching, 418, 426–427
 search, 250–253
 search engine, 4
 search engine optimization, 258
 search length, 349
 search session, 575
 second price auction model, 591
 seed, 142
 segmentation, 134
 selected completely at random, 200
 selective query expansion, 346
 self-training, 178–179
 semantic orientation, 463, 473
 semantic similarity, 337
 semi-supervised clustering, 171
 semi-supervised learning, 141
 sensitivity, 83
 sentence-level sentiment
 classification, 474
 sentiment analysis, 459–514
 sentiment classification, 469–474
 sentiment consistency, 478
 sentiment orientation, 463
 sentiment words, 470, 477
 separate-and-conquer, 93
 sequence, 42
 sequential covering, 87, 373
 sequential crawler, 312
 sequential pattern, 550
 sequential pattern mining, 6, 43–56
 frequent sequence, 42
 GSP, 43–45
 k-sequence, 42
 minimum support, 42
 MS-PS, 52–53
 multiple minimum supports,
 45–49, 52–53
 minimum item support, 45
 PrefixSpan, 49–51
 sequence, 42
 sequential pattern, 42
 contain, 42
 element, 42
 itemset, 42
 k-sequence, 42
 length, 42
 sequence, 42
 size, 42
 subsequence, 42
 supersequence, 42
 support, 42
 sequential rule, 54–56
 class sequential rule, 55–56
 label sequential rule, 54–55
 server log, 530
 session, 532, 536, 577
 session identification, 529
 session length, 575
 sessionization, 536–538
 set expansion, 186, 499
 set instance, 368
 set type, 367
 shingle method, 231–232
 sibling locality, 335
 similarity function, 124
 similarity group, 133
 simple domain, 432
 simple matching distance, 153
 simple tree matching, 387–389
 single word query, 252
 single-link method, 149
 singular value, 244
 singular value decomposition,
 243–246, 565–571
 skewed class distribution, 79
 small-world, 357
 smoothing, 103, 107, 220
 social choice theory, 255
 social media, 459
 social network analysis, 9, 269–303
 soft-margin SVM, 118
 spam hosts, 584–585
 spam queries, 584–585
 spamming, 212, 257–263
 sparse region, 163

- sparseness, 23
- specificity, 83
- spectral clustering, 166
- spectral graph transducer, 181
- spider, 311
- spider trap, 320–321
- sponsored Ads, 572
- sponsored results, 585
- sponsored search, 572, 589
- sponsored search advertising, 585–586, 590
- spy technique, 191–192
- squared Euclidean distance, 152
- SSE, *see* sum of squared error
- standard gradient descent, 568, 571
- standardization of words, 428
- start rule, 371–372
- stationary probability distribution, 281
- statistical language model, 219–220
- stem, 228
- stemmer, 228
- stemming, 228, 318, 428
- STM, *see* simple tree matching
- stochastic gradient descent, 568
- stochastic matrix, 280, 281, 284
- stopword removal, 214, 227, 280, 428
- string matching, 384–386
- strongest rule, 97–99
- strongly connected, 283
- structure data, 532
- structured data extraction, 363–419
- structured summary, 467
- subjectivity, 466
- subjectivity classification, 466, 474
- subsequence, 42
- subspace clustering, 166
- sufficient match, 381
- sum of squared error, 137
- summary of opinions, 460
- superlative comparisons, 494
- superlatives, 493
- supersequence, 42
- supervised approach, 509–511
- supervised learning, 6, 63–127
 - assumption, 66
 - class attribute, 63
 - class label, 63, 109
 - classification function, 64
 - classification based on
 - associations, *see* classification
 - based on associations
 - decision tree, *see* decision tree
 - example, 63
 - instance, 63
 - k -nearest neighbor, *see* k -nearest neighbor classification
 - learning process, 66
 - model, 65
 - testing phase, 66
 - training data, 65
 - training set, 65
 - training phase, 66
 - unseen data, 65
 - test data, 65
 - naïve Bayesian, *see* naïve Bayesian classification
 - prediction function, 64
 - rule induction, *see* rule induction
 - SVM, *see* support vector machines
 - vector, 63
- support, 18, 42, 54
- support count, 18, 69
- support difference constraint, 30, 49, 52
- support vector, 115
- support vector machines, 109–123
 - bias, 109
 - complementarity condition, 114, 118
 - decision boundary, 110, 116
 - decision surface, 110
 - dual variable, 116
 - dual, 115
 - input vector, 109
 - input space, 120
 - kernel, 120–123
 - feature space, 120
 - Gaussian RBF kernel, 123
 - input space, 120
 - kernel function, 122–123
 - kernel trick, 123
 - polynomial kernel, 122–123

Kuhn-Tucker conditions, 114, 118
 Lagrange multiplier, 113, 118
 Lagrangian, 113
 linear learning system, 109
 linear separable case, 111–116
 linear non-separable case, 117–120
 margin hyperplane, 111–112
 margin, 111
 maximal margin hyperplane, 111, 116
 nonlinear SVM, 120–123
 normal vector, 111
 polynomial kernel, 123
 primal, 115
 primal Lagrangian, 115
 primal variable, 115, 119
 slack variable, 117
 soft-margin SVM, 118
 support vector, 115
 weight vector, 109
 Wolfe dual, 116
 surface Web, 438
 SVD, *see* singular value decomposition
 symmetric attribute, 153

T

tag tree, *see* DOM tree
 targeted advertising, 590
 TCP/IP, 3
 template, 364, 379
 term, 211, 213, 215
 term ambiguity, 572
 term frequency, 217
 term spamming, 258
 term-pageview matrix, 542
 test data, 65
 test set, 79
 testing phase, 66
 text clustering, 154
 text mining, 6
 TF, 337
 TF-IDF, 217, 337
 theme, 295–296
 Tim Berners-Lee, 2
 time extraction, 498

Timed PageRank, 286
 token, 370
 top N candidate, 437–438
 topic drift, 293
 topical crawler, 311, 330–347
 adaptive topical crawler, 341
 best-first variation, 338–341
 best-N-first, 330, 340
 Clever, 340
 cluster hypothesis, 333
 InfoSpiders, 340, 344
 lexical topology, 332
 link topology, 333
 linkage locality, 336
 link-cluster conjecture, 333
 link-content conjecture, 333
 reinforcement learning, 343–346
 sibling locality, 335
 topology refinement, 376
 training data, 65, 79
 training phase, 66
 training set, *see* training data
 transaction, 17
 transaction data, 527
 transaction matrix, 541
 transduction, 179
 transductive Support Vector
 Machines, 179–180
 transductive SVM, *see* transductive
 Support Vector Machines
 transitive property, 437
 tree matching, 231, 384, 386
 simple tree matching, 387–388
 normalized tree matching, 389
 tree pruning, 76
 true negative, 81
 true negative rate, 83
 true positive, 81
 true positive rate, 83
 tuple instance, 368
 tuple type, 367

U

unary coding, 238
 union-free regular expression, 383, 412

universal crawler, 10, 311, 323
 unlabeled examples, 172–203
 unordered categorical, 152
 unsupervised learning, 65, 133–166

- cluster, 133
- cluster representation, 144–147
- clustering, 133–165
- cluster evaluation, *see* cluster evaluation
- data standardization, *see* data standardization
- distance function, *see* distance function
- hierarchical clustering, *see* agglomerative clustering
- k*-means clustering, *see k*-means clustering
- mixed attributes, 157–159
- partition, 134, 135
- segmentation, 134

 URL, 2
 usage data, 7, 530
 usage-based clustering, 545
 user activity record, 535
 user data, 532
 user identification, 534
 user intent, 572
 user profile, 527, 556–558
 user transactions, 540
 user-agent, 261, 353
 user-based recommendations, 559
 user-pageview matrix, 541
 utility of reviews, 514–515

V

valence shifters, 482
 validation set, 78, 80
 variable-byte coding, 242
 vector space model, 216–219

- cosine similarity, 218
- IDF, *see* inverse document frequency
- Okapi, 218
- inverse document frequency, 217
- normalized term frequency, 217

pivoted normalized weighting, 219
 term frequency, 217
 TF, *see* term frequency
 TF-IDF scheme, 217
 vocabulary, 215
 view graph, 580
 virtual society, 1
 visitor segmentation, 545
 visual information, 396, 406
 vocabulary search, 234

W

W3C, *see* World Wide Web Consortium
 Ward's method, 150
 Web, 1, 2

- CERN, 2
- distributed hypertext system, 2
- history, 2
- HTML, 2,
- HTTP, 2
- HyperText Markup Language, 2
- HyperText Transfer Protocol, 2
- Tim Berners-Lee, 2
- URL, 2

 Web content mining, 7
 Web data model, 366–368

- basic type, 367
- flat relation, 367
- flat set type, 367
- flat tuple type, 367
- instance, 367
- list, 369
- nested relation, 366–368
- set instance, 368
- set node, 367
- set type, 367
- tuple instance, 368
- tuple node, 367
- tuple type, 367

 Web database, 425
 Web mining, 6
 Web mining process, 7
 Web page pre-processing, 229
 Web query interface, 425, 438–454

- clustering based approach, 441–444
- correlation based approach, 440–447
- deep Web, 438
- global query interface, 425, 449
- instance based approach, 447–449
- inter-site schema matching, 449
- intra-site schema matching, 449
- label, 439
- name, 439
- schema model, 439
- surface Web, 438
- Web search, 250
- Web server access logs, 530
- Web spam, 257–263
 - combating spam, 262–263
 - content spamming, 258
 - content hiding, 261
 - cloaking, 261
 - directory cloning, 260
 - in-link spamming, 260
 - link spamming, 259
 - out-link spamming, 260
 - redirection, 261
 - term spamming, 258
 - search engine optimization, 258
 - user-agent field, 261
- Web structure mining, 7
- Web usage mining, 7, 527–593
- Weighted Euclidean distance, 152
- within-k-neighborhood, 580
- World Wide Web, 1
- World Wide Web Consortium, 4
- WorldWideWeb, 2
- wrapper generation, 363–420
 - building DOM tree, 396–397
 - center star method, 390
 - center string, 390
 - conflict resolution, 405
 - data record, 363, 368, 404
 - data region, 364, 398, 404
 - DeLa, 420
 - DEPTA, 420
 - disjunction or optional, 416–417
 - EXALG, 420
 - extraction based on a single list page, 397–413
 - extraction based on multiple pages, 413–415
 - generalized node, 400–401
 - grammar induction, 409
 - HTML code cleaning, 396
 - IEPAD, 420
 - MDR, 402, 420
 - multiple alignment, 390–391
 - nested data record, 407–413
 - NET, 407
 - node pattern, 409
 - partial tree alignment, 391–396
 - regular expression, 382, 409–411, 415–416
 - RoadRunner, 414–415
 - seed tree, 392
 - simple tree matching, 387–388
 - STM, *see* simple tree matching
 - string edit distance, 384–386
 - tree matching, 386–389
 - tidy, 396
 - union-free regular expression, 383, 412, 414
 - visual information, 396, 406
- wrapper induction, 370–381
 - active learning, 377
 - co-testing, 377
 - end rule, 371–372
 - informative example, 377
 - instance-based wrapper learning, 378–381
 - landmark, 371
 - list iteration rule, 371
 - perfect disjunct, 373
 - rule learning, 373–377
 - sequential covering, 373
 - start rule, 371–372
 - token, 370
 - wrapper maintenance, 378
 - wrapper repair, 378
 - wrapper verification, 378
- wrapper repair problem, 378
- wrapper verification problem, 378
- WWW conference, 4

Y

Yahoo!, 4

Z

zero count, 102

Zipf law, 574

z-score, 155–156