

# Index

- ACE-1, 164
- ACE-2
  - annotations, 164, 165
  - evaluation, 164–166
- acquisition bottleneck, 64
- activity networks, 198
- AdaBoost algorithm, 77, 78, 120
- agglomerative algorithms, 85
- Agrawal, C.C., 9, 24
- AI tasks, 64
- algorithm(s). *See also* Apriori algorithm; Borders
  - algorithm
  - (LP)<sup>2</sup>, 120
  - 3-D rendering, 219
  - AdaBoost, 77, 78, 120
  - agglomerative, 85
  - association generating, 26
  - BASILISK, 173–174
  - bootstrapping, 171
  - brute force, 85
  - Buckshot, 86, 88
  - BWI, 119–120
  - classic graph analysis, 260
  - clustering, 85–88
  - convex optimization, 72
  - cores and, 258–259
  - covering, 121
  - CRF, 121
  - Delta, 36, 41
  - documents structured by, 57
  - EM, 78, 90–91
  - EM-based mixture resolving, 85, 87
  - episode based, 41
  - evaluating, 121
  - FACT's, 49
  - force-directed graph layout, 245, 246
  - forward-backward, 134, 141
  - frequent concept set, 24, 25
  - FUP, 36, 41
  - FUP<sub>2</sub>, 36, 41
  - general graphs fast, 248
  - HAC, 85, 87–88
  - HMM, 121
  - Hobbs, 112
  - human-language processing, 60
  - IE, 98, 119
  - incremental, 30, 36
  - inductive, 119, 121
  - ISO-DATA, 86
  - KK, 247
  - K-means, 85, 86, 88
  - knowledge discovery, 2, 17, 193
  - layout, 245, 246
  - learning, 68
  - MEMM, 121
  - metabootstrapping, 170
  - mixture-resolving, 87
  - ML, 70
  - Naive, 112
  - optimization, 246
  - O-Tree, 124–125
  - pattern-discovery, 1, 5
  - preprocessing methodologies, 57
  - probabilistic extraction, 121
  - Ripper, 74, 298
  - salience, 114
  - search, 36, 178, 236
  - sequential patterns mining, 30
  - shuffling, 85
  - SOMs generated, 213, 216–217
  - spring-embedder, 245
  - SVM, 76–77
  - tasks for, 58
  - text mining, 5, 8
  - Viterbi, 133–134, 138, 141
  - WHISK, 119

- alone, maximal associations and, 27
- ambiguities
  - part-of-speech, 58
- analysis
  - banking, 280
  - critical path, 198, 235
  - data, 64
  - dependency, 61
  - domain, 105
  - Industry Analyzer corporate, 292–294
  - lexical, 105, 106, 107
  - linguistic, 109
  - morphological, 59, 105
  - patent, 295, 298
  - sentence, 109
  - syntactic, 105
  - textual, 146–152
  - time-based, 30
  - trend, 9, 30–31, 41, 299, 303
- anaphora
  - NLP, 118
  - one-, 111
  - ordinal, 111
  - pronominal, 110
- anaphora resolution, 109–119
  - approaches to, 109, 112, 113–114, 116, 117–119
- annotations
  - ACE-2, 164, 165
  - corpus, 109, 166
- answer-sets, 1, 23
- antecedent
  - closest, 118
  - most confident, 118
  - nonpronominal preceding, 118
- application. *See also* Document Explorer
  - application; GeneWays; knowledge discovery in text language, application; Patent Researcher; text mining
  - applications
  - area, 202
  - business intelligence, 280
  - creating custom, 285
  - horizontal, 295
  - KDD, 13
  - patent analysis, 295
  - TC, 64, 65–66
  - text mining, xi, 8
- apposition, 110
- Apriori algorithm, 24, 36
  - associations generated with, 37
  - textual application of, 39
- architects. *See* system, architects
- architecture
  - considerations, 192–194
  - FACT's, 46–47
  - functional, 13, 192
  - GeneWays', 308–310
  - IE, 104–109
  - Industry Analyzer, 281–288
  - open-ended, 116
  - preprocessing, 58
  - system, 46–47, 186
  - text mining system's, 13–18
- articulation points, 260
- assignment function *F*, 66
- association(s), 19, 25–26. *See also* ephemeral
  - associations; maximal associations
  - algorithm for generating, 26
  - Apriori algorithm generation of, 37
  - browsing tools for, 181
  - clustering, 181
  - concept, 9
  - concept sets and, 25
  - constraints, 181–183
  - discovery of, 24, 45, 46
  - displaying/exploring, 180–182
  - ephemeral, 30, 32
  - generating, 40
  - graphs, 198–200
  - left-hand side (LHS) of, 26, 45, 181
  - M-, 28
  - market basket type, 24, 25, 39
  - overwhelming, 181
  - partial ordering of, 202
  - query, 45, 46
  - right-hand side (RHS) of, 26, 45, 200
- association rules, 24, 25, 27, 182
  - circle graphs and, 208
  - definitions for, 40
  - discovering, 26–27
  - modeling, 210
  - search results, 36
  - for sets, 200
- attribute(s)
  - extracting, 96
  - relationship rules, 42
- auditing environment, 94
- automata. *See* finite-state automata
- automated categorization, 67
- AutoSlog-TS system, 166–168
- average concept
  - distribution, 22, 23
  - proportion, 22
- background
  - constraints, 186
  - states, 149
- background knowledge, 8, 42, 274
  - access to, 8
  - concept synonymy and, 45
  - constraints crafted by, 45
  - creating, 16
  - document collections and, 45
  - domains and, 8
  - FACT's exploitation of, 46

- forms of, 42
- generalized v. specialized, 274–276
- GeneWays' sources of, 308
- Industry Analyzer implementation of, 281
- integrating, 45
- large amounts of, 275
- leveraging, 8
- maintenance requirement, 276
- pattern abundance limited by, 45
- polysemy and, 45
- preservation of, 16
- sources of, 275
- specialized, 275
- text mining systems and, 8, 16, 42, 44
- back-propagation, 75
- backward variable, 133, 141
- Bacon, Kevin. *See* Kevin Bacon game
- bagging, 77–78
- bag-of-words model, 68, 89
- banking analysts, 280
- baseline classifiers, 80
- BASILISK algorithm. *See* Bootstrapping Approach to Semantic Lexicon Induction using Semantic Knowledge
- Baum–Welsh reestimation formulas, 135, 136, 147, 151
- Bayesian approximation, 120
- Bayesian logistic regression (BLR), 71–72
- benchmark collections, 79–80
- best-first clustering, 118
- betweenness centrality, 252–253, 256
- bigrams, 5
- binary
  - categorization, 67
  - matrix, 243
  - predicates, 16
  - relation, 242, 243
  - SVM classifiers, 76
  - tree, 73
- binary-valued trigger function, 139
- bins, 66
- BioGen Idec Inc., 283, 291
- biological pathways
  - information, 274
  - text mining,
- biological pathways information, 274
- biotech industry, 288–289
- BioWisdom company, 275
- BioWorld*, xi, 41, 281
- BioWorld Online*, 294
- block modeling, 262–266, 270
  - hijacker network and, 266–270
  - pajek, 268
- BLR. *See* Bayesian logistic regression
- Blum, A., 172
- Bonacich, P., 254
- Boolean constraints, 256
- Boolean expressions, 179
- boosted wrapper induction (BWI) algorithm, 119–120
- boosting, 77–78
  - classifiers, 77
  - indicators, 115
- boosting classifiers, 77
- boosting indicators, 115
- bootstrapping
  - algorithm, 171
  - categorization, 174–175
  - IE and, 166
  - introduction to, 166–168
  - meta-, 169
  - multi-class, 174
  - mutual, 168
  - problems, 172
  - single-category, 174
- Bootstrapping Approach to Semantic Lexicon Induction using Semantic Knowledge (BASILISK) algorithm, 173–174
- border sets, 36
- Borders algorithm, 36
  - benefits of, 37
  - notational elements of, 37
  - Property 1 of, 37
  - Property 2 of, 37
  - Stage 1 of, 37
  - Stage 2 of, 38
- Borgatti, S.P., 262
- Brown Corpus tag set, 60
- Brown, R.D., 228
- browsers, 177. *See also* Title Browser
  - character-based, 191
  - distribution, 238
  - Document Explorer, 238
  - interactive distribution, 238
- browsing. *See also* scatter/gather browsing
  - method
    - defined, 177–185
    - distributions, 179
    - hierarchical, 23
    - interface, 179, 189
    - interfaces, 276
    - methods, 179
    - navigational, 10
    - pattern, 14
    - result-sets for, 11
    - software for, 177
    - support operations, 203
    - text mining system, 10
    - tools, 181
    - tree, 15
    - user, 10, 13
- brute force
  - algorithm, 85
  - search, 9
- Buckshot algorithm, 86, 88

- business
  - intelligence, 279, 280
  - sector, 280
- BWL. *See* boosted wrapper induction
- C4.5 procedure, 73
- Cardie, Claire, 118
- Carnegie group, 70
- CART procedure, 73
- categorization. *See also* text categorization
  - attributes, 45
  - automated, 67
  - binary, 67
  - bootstrapping, 174–175
  - category-pivoted, 67
  - document-pivoted, 67
  - hard, 67
  - hierarchical Web page, 66
  - manual-based, 6, 12
  - methodologies of, 6
  - multilabel, 67
  - online, 67
  - patent analysis and, 298
  - POS tag set, 60
  - POS word, 60
  - preprocessing methodology, 57
  - problems, 82
  - relationship based, 45
  - rule-based, 6
  - single-label, 67
  - soft (ranking), 67
  - systems, 91
- categorization status value (CSV), 67
- category connecting maps, 211–212, 239
- category domain knowledge, 42
- CDM-based methodologies, 7, 12
- centrality, 249
  - betweenness, 252–253, 256
  - closeness, 251
  - definitions used for, 249
  - degree, 249–251, 255
  - eigenvector, 253–254
  - measures of, 249
  - natural language text, 1
  - power, 254–255
- centralization, network, 255–256
- centroids, medoids v., 90
- character(s), 5, 8
  - classes,
  - representations of, 5
- character-level regular expressions,
- chi-square measures, 69, 200
- chunking, NP, 154
- CIA World Factbook, 46, 48, 50
- circle graphs, 190, 208–213, 286
  - click-sensitive jumping points of, 210
  - controls of, 211
  - data modeling by, 213
  - interactivity, 190
  - mouse-overs and, 210
  - multiple, 212–213
  - nodes, 292
  - style elements of, 210
  - usefulness of, 208
  - visualization, 292
- classes
  - character,
  - equivalence, 201–202
- classification
  - line, 153
  - schemes, 131
- classifier(s)
  - baseline, 80
  - binary SVM, 76
  - boosting, 77
  - building, 66
  - common, 68
  - comparing, 80
  - continuous, 67
  - decision rule, 73–74
  - decision tree, 72–73
  - example-based, 75–76
  - ith*, 77
  - k* different, 77
  - kNN, 75
  - machine learning, 70
  - ME, 153
  - NB, 71, 78, 90–91
  - probabilistic, 71, 78
  - Rocchio, 74–75
  - stateless ME, 153
  - symbolic, 72
  - text, 76, 79–80
  - training, 79
- classifier committees, 77–78
- ClearForest Corporation, 294
- ClearForest Text Analytics Suite, 294, 296
- ClearResearch*, 231
- closeness centrality, 251
- cluster(s)
  - chain-like, 87
  - complete-link, 87, 88
  - gathering, 83
  - k*, 88
  - labels, 91
  - postprocessing of, 86
  - scattering, 83
  - single link, 87, 88
- cluster hypothesis, 82
- cluster-based retrieval, 84
- clustering. *See also* nearest neighbor clustering
  - algorithms, 85–88
  - associations, 181
  - best-first, 118
  - defined, 70, 82
  - disjoint, 75

- documents grouped by, 83
- flat (partial), 85
- good, 84
- hard, 85
- hierarchical, 83
- optimization, 85
- problem, 84–85, 89
- quality function, 84, 92
- query specific, 83
- soft, 85
- tasks, 82–84
- term, 69
- text, xi, 89, 91–92
- tools, 11, 184–185
- unsupervised, 185
- usefulness of, 82
- users and, 83
- of vertices, 264
- CO. *See* coreference task
- CogNIAC, 113
- collections, benchmark, 79–80
- color
  - assigning of, 279
  - coding, 45, 289
  - GUI palette of, 279
- Columbia University, 307, 310–311
- column-orthonormal, 90
- combination graphs, 212–213
- command-line query interpreters, 10
- committees
  - building, 77
  - classifier, 77–78
- components
  - bi-connected, 260
  - presentation layer, 14
  - strong, 260
  - weak, 260
- computational linguistics, 1, 3
- concept(s), 5–8
  - associations of, 9
  - context, 33
  - co-occurrence, 9, 23
  - DIAL language,
  - distribution, 21
  - distribution distance, 29
  - documents, 23
  - extraction, 7
  - features, 12
  - graphs, 202–204
  - guards, 328–329
  - hierarchy node, 20
  - identifiers, 6, 197
  - interdocument association, 9
  - keywords v., 12
  - link analysis, 226
  - names, 326
  - occurrence, 19
  - output, 156
  - patterns, 10
  - proportion, 22, 29
  - proportion distance, 29
  - proportion distribution, 21, 22
  - representations, 7
  - selection, 19, 20
  - sentences, 321
  - subsets of, 201
  - synonymy, 45
- concept hierarchies, 43
  - editing tools, 184
  - maintaining, 183
  - navigation/exploration by, 182
  - node, 20
  - roles of, 182
  - taxonomy editors and, 183–184
- concept set(s), 22
  - associations and, 25
  - cosine similarity of, 201
  - display of, 196
  - graphs, 196, 197
- concision, 191
- conditional models, 140
- conditional probability, 71, 142
  - computing, 143
- conditional random fields (CRFs), 142–144
  - algorithm, 121
  - chunk tagger, 155
  - chunker, 154
  - development of, 153
  - formalism, 153
  - linear chain, 142
  - part-of-speech tagging with, 153–154
  - problems relating to, 143
  - shallow parsing with, 154–155
  - textual analysis and, 153–155
  - training of, 144
- conditions, 120
- confidence, 25
  - M-, 27, 28
  - threshold, 181
- constants. *See also* string, constants
  - Boolean, 328
  - rule, 327–328
- constituency grammars, 60–61
- constraint(s), 42
  - accessing, 185–186
  - association, 181–183
  - background, 186
  - background knowledge crafting of, 45
  - comparison, 328
  - controls, 191
  - FACT's exploitation of, 46
  - functions, 139
  - leveraging, 276
  - logic of, 186
  - parameters, 45
  - Patent Researcher's, 298–299

- constraint(s) (*cont.*)
  - quality, 186
  - query, 278
  - redundancy, 186
  - refinement, 11, 14, 19–41, 191, 284–285, 298–299
  - search, 178, 203
  - syntactical, 186
  - types of, 186
- CONSTRUE system, 70, 73
- contained matches, 101
- context. *See also* temporal context relationships
  - concept, 33
  - DIAL language,
  - focus with, 191
  - phrase, 33
  - relationships, 32, 33
- context equivalence, 202
- context graphs, 30, 32, 33–35
  - components of, 33
  - defined, 33
- context-dependent probabilities, 149, 152
- continuous real-valued functions, 74
- control elements, 191
- controlled vocabulary, 65
- convex optimization algorithms, 72
- co-occurrence
  - concept, 9, 23
  - frequency of, 24
  - relationships, 12
- core
  - algorithm for finding, 258–259
  - vertices of, 258
- core text mining operations, 14, 19–41, 284–285
  - Patent Researcher and, 298–299
- coreference
  - function–value, 111
  - part–whole, 112
  - proper names, 110
  - resolution, 109, 112
- coreference task (CO), 99
- coreferring phrases, 109
- corporate finance, 273
  - business intelligence performed in, 279
  - text mining applications, 284
- corpus
  - annotated, 109, 166
  - MUC-4, 170
- cosine similarity, 90, 200, 201
- Costner, Kevin, 248
- cotraining, 78, 172
- cover equivalence, 202
- covering algorithm, 121
- $\sigma$ -cover sets, 24. *See also* singleton,  $\sigma$ -covers
  - FACT's generation of, 49
- CRFs. *See* conditional random fields
- critical path, 234
  - analysis, 198, 235
  - diagrams, 234
  - graphs, 235
- Croft, W.B., 76
- cross-referencing, 6
- CSV. *See* categorization status value
- Cui, Y., 197, 198
- CUTenet, 310
- Cutting, D.R., 92
- cycle, 243
  - graph, 248–249
  - transmission/emission, 131
- DAGs. *See* directed acyclic graphs
- Daisy Analysis, 225
- Daisy Chart, 225
- DAML, 275
- DARPA, 96
- data
  - abstraction, 91
  - analyzing complex, 64
  - Apriori algorithm and textual, 39
  - clustering, 14, 88–92
  - color-coding of, 45
  - comparing, 29
  - currency, 36
  - discovering trends in textual, 30
  - dynamically updated, 36
  - exploration, 184–185
  - GeneWays' sources of, 308
  - identifying trends in, 9
  - inter-document's relationships with, 2
  - modeling, 213
  - Patent Researcher, 297
  - patterns of textual, 40
  - preparing, 57
  - scrubbing/normalization of, 1
  - sparseness, 136–137, 148
  - textual, 88–92, 189
  - thresholds for incremental, 39
  - unlabeled, 78
  - unstructured, 194
  - visualization of, 217
- data mining
  - analysis derived from, 10
  - border sets in, 36
  - pattern-discovery algorithms, 1
  - preprocessing routines, 1
  - presentation-layer elements, 1
  - text mining v., 1, 11
  - visualization tools, 1
- database
  - GenBank, 308
  - MedLine, 11, 78, 275
  - OLDMEDLINE, 12
  - relational, 4
  - Swiss-Prot, 308
- decision
  - rule classifier, 73–74
  - tree (DT) classifiers, 72–73
- decomposition, singular value, 89–91
- definite noun phrases, 117

- definiteness, 115
- degree centrality, 249–251, 255
- Delta algorithms, 36, 41
- demonstrative noun phrases, 117
- dense network, 244
- dependency
  - analysis, 61
  - grammars, 61
- detection. *See* deviation, detection
- deviation
  - detection, 10, 13, 30, 32
  - sources of, 32, 41
- diagrams
  - critical path, 234
  - fisheye, 227–231
- DIAL language, 283, 297
  - code, 320–322
  - concept, 317
  - concept declaration, 317
  - context, 321
  - Discovery Module, 319
  - engines, 317
  - examples, 329–330, 331–332, 333–336
  - information extraction, 318–319
  - module, 318–319
  - plug-in, 321
  - scanner properties, 320
  - searches, 321
  - sentence concept, 321
  - text pattern, 317–318
  - text tokenization, 320
- dictionaries, 106
- dimension reduction, 69, 89
  - LSI with, 89
  - SVD and, 90
- dimensionality
  - document collection's high, 215
  - document reduction, 89
  - feature, 4, 12
- direct ephemeral association, 31
- directed acyclic graphs (DAGS), 43, 61, 197
  - activity networks and, 198
  - ontological application of, 197
  - visualization techniques based on, 198
- directed networks, 249, 260
- disambiguation, 156
- disconnected spring graphs, 234
- discovery
  - association rules, 26–27
  - of associations, 24, 45, 46
  - ephemeral associations, 10, 13
  - frequent concept sets, 24, 39, 40
  - methods, 24
- Discovery Module, 319
- disjoint clusters, 75
- dissimilarity matrix, 247
- distance, referential, 116
- distribution(s), 9, 19–23
  - average, 23
- average concept, 22
- Boolean expressions generation of, 179
- browsing, 179
- comparing specific, 23
- concept, 21
- concept co-occurrence, 23
- concept proportion, 21, 22
- conditional probability, 142
- interestingness and, 29–30
- patterns based on, 29, 32, 301
- queries, 205, 292
- text mining systems and, 22, 29
- topic, 31
- divide-and-conquer strategy, 58
- DNF rules, 73
- document(s)
  - algorithm's structuring of, 57
  - association of, 9
  - bag-of-words, 89
  - binary, 73
  - binary vector of features as, 4
  - bins, 66
  - clustering of, 83
  - collections of, 4
  - concept-labeled, 23
  - correlating data across, 2
  - co-training, 78
  - data relationships and, 2
  - defined, 2–4
  - dimensionality reduction of, 89
  - document collection's adding of, 36
  - features of, 4–8, 12
  - field extraction, 59, 146–148
  - free format, 3
  - good, 66
  - IE representation of, 95
  - irrelevant, 66
  - managing, 64
  - manually labeling, 78
  - meaning, 59
  - native feature space of, 5
  - natural language, 4
  - news feed, 32
  - O-Tree, 125
  - patent, 304
  - portraying meanings of, 5
  - proportion of set of, 20
  - prototypical, 3
  - quantities for analyzing, 20
  - relevant, 66
  - representations of, 4, 5, 6, 7, 58, 68
  - retrieval of, 179
  - scope of, 3
  - semistructured, 3–4
  - sorting, 65–66
  - sources of, 59
  - tagging, 94
  - task structuring of, 57
  - test sets of, 79

- document(s) (*cont.*)
  - text, 3
  - typographical elements of, 3
  - unlabeled, 78
  - weakly structured, 3–4
- document collection(s)
  - analyzing, 30
  - application area of, 202
  - background knowledge and, 45
  - defined, 2–3, 4
  - documents added to, 36
  - dynamic, 2
  - high-dimensionality, 215
  - Industry Analyzer, 281
  - processed, 15
  - PubMed as real-world, 2
  - scattering, 83
  - static, 2
  - subcollection, 19, 30
- Document Explorer application, 18
  - browsers, 238
  - development of, 235
  - knowledge discovery toolkit of, 238
  - modules, 236
  - pattern searches by, 236
  - term hierarchy editor, 237–238
  - visualization tools of, 236, 238
- domain(s), 8
  - analysis, 105
  - background knowledge and, 8
  - customization, 276
  - defined, 42
  - domain hierarchy with, 43
  - hierarchy, 43
  - knowledge, 42, 58, 59
  - lexicons, 44
  - ontology, 42–43, 44, 51
  - scope of, 8, 42
  - semistructured, 121
  - single application, 12
  - terminology preference, 116
  - text mining system's, 16
- DT classifier. *See* decision tree (DT) classifiers
- Eades, P., 231, 246
- edges, 33, 35
- Eigenvector centrality, 253–254
- EM. *See* expectation maximization
- e-mail, 3
- energy minimization, 248
- engineering
  - knowledge, 70, 155
- engines
  - DIAL, 317
  - GeneWays' parsing, 308
  - IE, 95
  - pronoun resolution, 113
  - query, 16
  - search, 82, 199
- entities
  - choosing query, 275
  - content-bearing, 94
  - equivalence between, 260–261
  - extracting, 96, 149, 150, 156, 164
  - hierarchy, 95
  - IE process and relevant, 95
  - links between, 242
  - multiple, 101
  - real world, 307
- ephemeral associations, 30
  - defined, 31
  - direct, 31
  - discovery, 10, 13
  - examples, 31
  - inverse, 32
- episodes, algorithms based on, 41
- equivalence
  - classes, 201–202
  - context, 202
  - cover, 202
  - entity, 260–261
  - first, 202
  - regular, 261
  - structural, 261
- Erdős number, 248
- Erdős, Paul, 248
- error(s)
  - false negative, 74
  - false positive, 74
  - matrix, 268
  - precision, 66
  - recall, 66
- non-Euclidean plane, 217
- evaluation workbench, 116
- event
  - example of, 94
  - extraction, 96
- Everett, M.G., 262
- exact matches, 101
- example-based classifiers, 75–76
- expectation maximization (EM), 78, 87, 90–91
  - mixture resolving algorithm, 85, 87
- Explora system, 18
- exploration
  - concept hierarchy, 182
- external ontologies, 100
- extraction. *See also* field extraction; Nymble
  - algorithms, 121
  - attribute, 96
  - concept, 7
  - DIAL examples of, 329–330, 331
  - domain-independent v. domain-dependent, 98
  - entities, 96, 149, 150, 156, 164
  - event, 96
  - fact, 96
  - feature, 69, 84, 283
  - grammars, 138
  - HMM field, 146–148



- information, 2, 11, 61–62, 119
- literature, 190
- Los Alamos II-type concept, 7
- relationship, 156, 164–166
- ST, 99
- structural, 122
- TEG, 164
- term, 6, 12, 95, 283
- text, 96
- TR, 99
- visual information, 122
- F* assignment function, 66
- FACT. *See* Finding Associations in Collections of Text
- facts, 94
  - extracting, 96
- FAQ file, 153
- FBI Web site, 244
- feature(s). *See also* native feature space
  - concept-level, 12
  - dimensionality, 4, 12
  - document, 4–8, 12
  - extraction, 69, 84, 283
  - linguistic, 100
  - markable, 117
  - orthographic, 100
  - relevance, 69
  - selection, 68–69, 84, 100
  - semantic, 100
  - space, 85
  - sparsity, 4
  - state, 142
  - synthetic, 69
  - transition, 142
- Feldman, R., 46
- field extraction, 59, 146
  - location, 149
  - speaker, 149, 152
- files
  - PDF, 3
  - word-processing, 3
- filters, 14
  - fish-eye view, 229–230
  - information, 14
  - personalized ad, 66
  - redundancy, 201–202
  - simple specification, 185–186
  - text, 65–66
- finance. *See* corporate finance
- Finding Associations in Collections of Text (FACT), 18, 46–51
  - algorithm of, 49
  - background knowledge exploitation by, 46
  - constraints exploited by, 46
  - $\sigma$ -cover sets generated by, 49
  - designers of, 50
  - implementing, 47–49
  - performance results of, 50–51
  - query language, 46
  - system architecture of, 46–47
- Findwhat, 199
- finite-state automata, 156
- first equivalence, 202
- fish-eye diagrams, 227–231
- fish-eye interface, 230
- fish-eye views
  - distorting, 228, 230
  - effectiveness of, 230–231
  - filtering, 229–230
- fixed thresholding, 67
- focus with context, 191
- force-directed graph layout algorithms, 245, 246
- formats, converting, 59
- formulas
  - Baum–Welsh reestimation, 135, 136
  - network centralization, 255
- forward variable, 132, 141
- forward–backward algorithm, 134, 141
- forward–backward procedure, 132–133
- FR method, 246–248
- fractal approaches, 229
- fragments, text, 109
- frames
  - hierarchy, 95
  - structured objects as, 95
- Freitag, D., 146
- frequent concept sets, 9, 23–24
  - algorithm for generating, 24, 25
  - Apriori-style, 36
  - discovery methods for, 24, 39, 40
  - generating, 24
  - identifying, 25
  - natural language in, 24
  - near, 25
  - $\sigma$ -cover sets as, 24
  - $\sigma$ -covers as, 24
- front-end, de-coupled/loosely coupled, 193
- Fruchterman, T., 232, 248. *See also* FR method
- function(s)
  - binary-valued trigger, 139
  - clustering quality, 84
  - constraint, 139
  - continuous real-valued, 74
  - similarity, 84, 200–201
  - trigger-constraint, 153
- functional architecture, 13, 192
- functionality
  - GeneWays', 308–310
  - Industry Analyzer system, 290
  - Patent Researcher's, 296–300
  - types of, 10
- function–value coreference, 111
- FUP algorithm, 36, 41
- FUP<sub>2</sub> algorithm, 36, 41
- Furnas, G., 228
- fuzzy search, 184

- game. *See* Kevin Bacon game
- Gaussian priors, 71
- Gelbukh, A., 9
- GenBank database, 308
- Gene Ontology Consortium, 43, 275
- Gene Ontology<sup>TM</sup> knowledge base, 43, 51, 197
- generalized iterative scaling, 140, 144
- generative models, 140
- generative process, 131
- generic noun phrases (GN), 171. *See also* noun phrases; phrases, coreferring; pronoun resolution engine; proper noun phrases
- GeneWays, xi, 307
- architecture/functionality of, 308–310
- background knowledge sources, 308
- core mining operations of, 310
- core mission of, 307
- CUtenet of, 310
- data sources, 308
- GUI, 310
- implementation/usage, 310
- Industry Analyzer comparison with, 307
- Parsing Engine, 308
- Patent Researcher comparison with, 307
- preprocessing operations, 308–310
- presentation layer elements, 310
- Relationship Learner module, 309
- specialized nuances of, 308
- Synonym/Homonym Resolver, 309
- GENomics Information Extraction System (GENIES), 308
- givenness, 115
- GN. *See* generic noun phrases
- Google, 200. *See also* search
- grammars. *See also* stochastic context-free grammars
- ambiguity of, 137
- canonical, 137
- constituency, 60–61
- dependency, 61
- extraction, 138
- nonstochastic, 137
- TEG, 157, 158
- graph(s). *See also* circle graphs; line graphs; singleton, vertex
- analysis algorithm, 260
- combination, 212–213
- concept, 202–204
- concept set, 196, 197
- connected spring, 234
- connection, 200
- context, 30, 32, 33–35
- critical path, 235
- cycles in, 248–249
- disconnected spring, 234
- drawing large, 248
- fast algorithm, 248
- general undirected, 247
- histogram-based trend, 307
- multivertex, 198
- node-and-edge, 227
- paths, 248–249
- simple concept, 195–205, 239, 286, 294
- simple concept association, 198–200
- spring embedded network, 231
- temporal context, 30, 32, 35
- theory, 242
- trend, 30, 32, 35, 239
- graphical user interface (GUI), 14, 46. *See also* interface
- developers, 226
- display modalities of, 178
- easy to use controls of, 177
- GeneWays, 310
- histogrammatic representations situated in, 205
- palette of colors, 279
- Patent Researcher's, 299–300
- queries, 284
- text mining application's, 177
- grouping, perceptual, 59, 123–124
- GUI. *See* graphical user interface
- HAC algorithm. *See* hierarchical agglomerative clustering algorithm
- Hadany, R., 248
- Harel, D., 248
- heuristics. *See* syntactic heuristics
- hidden Markov model algorithm, 121
- hidden Markov models (HMMs), 131–137. *See also* maximum entropy Markov model
- assumptions of, 132
- characteristics, 147
- classes of states, 147
- classic, 151
- defined, 131–137
- document field extraction by, 146–148
- entity extractor, 164
- field extraction, 146–148
- fully connected, 153
- MEMM outperformed by, 153
- Nymble and, 150
- optimal, 148
- POS tagging and, 156
- problems related to, 132
- single-state, 148
- textual analysis and, 146–152
- topology, 147, 148
- training, 135–136
- hierarchical agglomerative clustering (HAC) algorithm, 85, 87–88
- hierarchy. *See also* trees, hierarchical
- clustering, 83
- concept, 43
- domain, 43
- editor, 237–238

- entity/frame, 95
- internal nodes of, 23
- IS\_A-type, 282
- object's, 123
- ontology, 46
- shrinkage, 148
- Web page, 66
- hijacker(s)
  - network, 266–270
  - 9/11, 244
- histogram(s), 205–207, 286
  - distribution pattern demonstration in, 301
  - graphs, 307
  - interactivity of, 207
  - link analysis and, 226
- HMMs. *See* hidden Markov models
- Hobbs algorithm, 112
- holonymy, 112
- homonymy, 69
- Honkela, T., 216
- Hooke's Law, 246
- HSOM. *See* hyperbolic self-organizing map
- HTML
  - Web pages, 3, 66
  - WYSIWYG editor, 3
- human(s)
  - knowledge discovery view by, 13, 17, 177
  - language processing, 60
- hybrid system
  - introduction to, 155–156
  - TEG as, 156
- hybrid tools, 221–224
- hyperbolic non-Euclidean plane, 217
- hyperbolic self-organizing map (HSOM), 225
- hyperbolic trees, 217–219
- hypernyms, 184
- hyperplanes, 76
- hypertext, 66
- hyponyms, 184
- hypothesis
  - cluster, 82
  - weak, 77
- IBM, 199
- ID3 procedure, 73
- identical sets, 111
- identifiers, 6, 197
- IdentiFinder. *See* Nymble
- IE. *See* information extraction
- Imielinski, T., 24
- immediate reference, 115
- incremental algorithms, 30, 36
- incremental update schemes, 38
- indefinite noun phrases, 117
- indexing, 65, 83. *See also* latent semantic indexing
  - indexing
- indicating verbs, 115
- inductive algorithm, 119, 121
- inductive rule learning, 73
- Industry Analyzer system, 280
  - architecture/functionality of, 281–288
  - background knowledge implementation, 281
  - ClearForest Text Analytic's Suite and, 297
  - color coding, 289
  - core mining operations, 284–285
  - corporate analysis and, 292–294
  - document collection, 281
  - ease-of-use features of, 285
  - event-type query, 290
  - functionality, 290
  - GeneWays comparison with, 307
  - graphical menus, 288
  - implementation of, 282
  - merger activity with, 288
  - preprocessing operations, 282–284
  - presentation layer, 285–288
  - refinement constraints, 284–285
  - scope of, 280
  - search results, 291
  - term extraction, 283
  - visualization tools, 292
- inferencing, 108–109
- influence, 249
- information
  - age, x
  - biological pathways, 274
  - control of, 229
  - exploring, 292–294
  - filtering, 14
  - flow, 105–109
  - gain, 69
  - retrieval, 1, 2, 62, 82
- information extraction (IE), 2, 11, 61–62, 122
  - algorithms, 98, 119
  - architecture, 104–109
  - auditing environment, 94
  - benchmarks, 155
  - bootstrapping approach to, 166
  - DIAL language, 318–319
  - documents represented by, 95
  - engine, 95
  - evaluation, 100, 101
  - evolution of, 96–101
  - examples, 101, 102, 104
  - hybrid statistical, 155–166
  - information flow in, 105–109
  - knowledge-based, 155–166
  - MEMM for, 152–153
  - relevant entities and, 95
  - SCFG rules for, 155–166
  - schematic view of, 95
  - specialized dictionaries for, 106
  - statistical/rule-based, 156
  - structured, 122
  - symbolic rules of, 119
  - usefulness of, 94, 104

- input–output paradigm, 13
- inquiry, analytical, 273
- Inxight Software, 217
- interactivity
  - circle graph, 190
  - concept graph, 202–204
  - facilitating, 195
  - histogram, 207
  - user, 179, 189
- interestingness
  - defining, 29
  - distributions and, 29–30
  - knowledge discovery and, 40
  - measures of, 9, 179
  - proportions and, 29–30
- interface. *See also* graphical user interface
  - browsing, 179, 189, 276
  - fish-eye, 230
  - Patent Researcher's Taxonomy Chooser, 297
  - query language, 10
  - visualization, 191
  - WEBSOM's, 215
- interpreters, 10
- Iossifov, I., 310
- IS-A-type hierarchies, 282
- ISO-DATA algorithm, 86
- i*th classifier, 77
  
- Jones, R., 168, 169
  
- k* clusters, 88
- k* different classifiers, 77
- Kamada, T., 232, 248. *See also* KK method
- Kawai, S., 232, 248. *See also* KK method
- KDTL. *See* knowledge discovery in text language
- Kevin Bacon game, 248
- keywords
  - assigning, 65
  - concepts v., 12
- KK method, 246, 247
- K-means algorithm, 85, 86, 88
- k*-nearest neighbor (*k*NN) classifier, 75
- k*NN classifier. *See k*-nearest neighbor (*k*NN) classifier
- knowledge
  - base, 17
  - category domain, 42
  - distillation, 14
  - domain specific, 58, 59
  - engineering, 64, 70, 155
- knowledge discovery
  - algorithms, 2, 17, 193
  - distribution-type pattern's, 32
  - Document Explorer toolkit for, 238
  - human-centered view of, 13, 17, 177
  - interestingness and, 40
  - overabundance problems of, 179
  - Patent Researcher's, 299
  - patterns of, 14
  - problem-sets, 194
  - supplementing, 31
- knowledge discovery in text language (KDTL), 18, 52
  - application, 18, 236
  - queries, 52–54, 55, 236
- Kohonen maps, Kohonen networks. *See* self-organizing maps
- Kohonen, T., 213
- Koren, Y., 248
- Kuhn–Tucker theorem, 139, 140
  
- label
  - bias problem, 142
  - sequence, 144
- Lafferty, J., 153
- Lagrange multipliers, 139
- language. *See also* natural language processing; sublanguages
  - DIAL, 283
  - FACTS's query, 46
  - natural, 4
  - processing human, 60
  - query, 10, 14, 51–52, 177
  - soft mark-up, 3
- Laplace
  - priors, 71, 72
  - smoothing, 136
- Lappin, S., 113
- Larkey, L.S., 76
- latent semantic indexing (LSI), 69, 89
  - dimension reduction with, 89
- layout
  - algorithms, 245, 246
  - force-directed, 246
  - network, 244–248, 275
- learner, weak, 77
- learning. *See also* machine learning
  - algorithms, 68
  - inductive rule, 73
  - rules, 74
  - supervised, 70
- Leass, H.J., 113
- left-hand sides (LHSs), 26, 45, 181
- lemmas, 60
- lemmatization, 6, 283
- Lent, B., 9
- lexical analysis, 105, 106, 107
- lexical reiteration, 115
- lexicons, 8, 42, 44
  - domain, 44
  - external, 283
  - GN, 171
  - PNP, 171, 172
  - semantic, 169, 170

- LHSs. *See* left-hand sides
- libraries. *See also* National Library of Medicine  
graphing software, 207  
integration/customization of, 207
- life sciences, 273  
business sector, 280  
research,
- LINDI project, 18
- line graphs, 207–208  
link analysis and, 226  
multi-, 208  
as prototyping tools, 207
- linear least-square fit (LLSF), 74
- linguistic(s)  
computational, 1, 3  
features, 100  
processing, 283  
sentence analysis, 109
- link(s)  
detection, 230–231  
between entities, 242  
operations, 203–204
- link analysis, 225  
concepts, 226  
histograms and, 226  
line graphs and, 226  
software packages for, 271–272
- literature, extraction of, 190
- LLSF. *See* linear least-square fit
- Los Alamos II-type concept extraction, 7
- loss ratio parameter, 74
- Louis-Dreyfus, Julia, 248
- (LP)<sup>2</sup> algorithm, 120
- LSI. *See* latent semantic indexing
- Lycos, 199
- machine learning (ML), 64, 70–78, 166  
algorithms, 70  
anaphora resolution and, 117–119  
classifier, 70  
techniques, 70–71
- MacKechnie, Keith, 248
- mapping, structural, 125–127
- maps, category connecting, 211–212, 239
- marginal probability, 71
- markables, 117
- market basket  
associations, 24, 25, 39  
problems, 25, 39
- M-association, 28
- matches  
contained, 101  
exact, 101  
overlapped, 101
- matrix  
binary, 243  
dissimilarity, 247
- error, 268  
transmission, 143
- maximal associations. *See also* M-association;  
M-confidence; M-frequent; M-support  
alone and, 27  
M-factor of, 40  
rules, 27, 40
- maximal entropy (ME), 131, 138–140, 153
- maximal entropy Markov model (MEMM),  
140–141  
algorithm, 121  
comparing, 153  
HMM and, 153  
information extraction and, 152–153  
training, 141
- maximum likelihood estimation, 91
- McCallum, A., 146
- M-confidence, 27, 28
- ME. *See* maximal entropy
- measures  
centrality, 249  
chi-square, 69, 200  
interestingness, 9, 179  
network centralization, 255  
performance, 79  
similarity, 85  
uniformity, 152
- MedLEE medical NLP system, 309
- MedLine, 11, 78, 275
- medoids, centroids v., 90
- MEMM. *See* maximal entropy Markov  
model
- merger activity, 288–289
- meronymy, 112
- MeSH, 275
- MESH thesaurus, 65
- Message Understanding Conferences (MUC),  
96–101
- metabootstrapping, 169, 170
- methodologies. *See also* preprocessing  
methodologies  
categorization, 6  
CDM-based, 7, 12  
information extraction, 11  
term-extraction, 6, 12, 95, 283  
text mining, 9
- M-factor, 40
- M-frequent, 28
- Microsoft, 199
- middle-tier, 193
- Miller, James, 225
- minconf thresholds, 26, 40
- minimal spanning tree (MST), 88
- minimization, 248
- minsup thresholds, 26, 40
- Mitchell, T.M., 172
- mixture-resolving algorithms, 87
- ML. *See* machine learning

- models
  - block, 262–266, 270
  - conditional v. generative, 140
  - data, 213
  - Document Explorer, 236
  - ME, 138–140
  - probabilistic, 131
- module. *See also* Discovery Module
  - DIAL language, 318–319
  - Document Explorer application, 236
  - GeneWays Relationship Learner, 309
- Montes-y-Gomez, M., 8–10
- morphological analysis, 59, 105
- most probable label sequence, 144
- MSN, 199
- MST. *See* minimal spanning tree
- M-support, 27, 28
- MUC. *See* Message Understanding Conferences
- MUC-4 corpus, 170
- MUC-7 Corpus Evaluation, 164
- multilabel categorization, 67
- Murphy, Joseph, 307
- multivertex graphs, 198
- Mutton, Paul, 231
  
- Naive algorithm, 112
- Naive Bayes (NB) classifiers, 71, 78, 90–91
- named entity recognition (NER), 96, 164
- names
  - concept, 326
  - identifying proper, 106–107
  - proper, 97
  - thesaurus, 325
  - wordclass, 324–325
- NASA space thesaurus, 65
- National Cancer Institute (NCI), 282
- National Cancer Institute (NCI) Metathesaurus, 294
- National Center for Biotechnology Information (NCBI), 11
- National Institute of Health (NIH), 11
- National Library of Medicine (NLM), 2, 11, 275, 282
- native feature space, 12
- natural language processing (NLP), 4
  - anaphoric, 118
  - components, 60
  - elements, 117
  - field extraction and, 146
  - frequent sets in, 24
  - general purpose, 58, 59–61
  - MedLEE medical, 309
  - techniques, 58
- natural language text, 1
- navigation, concept hierarchy, 182
- NB classifiers. *See* Naive Bayes classifiers
- NCBI. *See* National Center for Biotechnology Information
- NCI. *See* National Cancer Institute
- near frequent concept sets, 25
- nearest neighbor clustering, 88
- negative borders, 36
- NER. *See* named entity recognition
- NetMap, 209
- NetMiner software, 272
- network(s). *See also* spring embedding, network
  - graphs
    - activity, 198
    - automatic layout of, 244–248, 275
    - centralization, 249, 255–256
    - clique, 244
    - complex, 75
    - dense, 244
    - directed, 249, 260
    - formulas, 255
    - hijacker, 266–270
    - layered display of, 259
    - layout, 244–248, 275
    - neural, 75
    - nonlinear, 75
    - partitioning of, 257–270
    - pattern matching in, 270
    - patterns, 242
    - self-loops in, 244
    - social, 242
    - sparse, 244
    - two-mode, 244
    - undirected, 260
    - weakness of, 260
- neural networks (NN), 75
- Ng, Vincent, 118
- ngrams, 156
  - construction of, 157
  - featureset declaration, 163
  - parent, 161
  - restriction clause in, 163
  - shrinkage, 163
  - statistics for, 159
  - token generation by, 159, 161
- NIH. *See* National Institute of Health
- 9/11 hijacker example, 244
- NLM. *See* National Library of Medicine
- NLP. *See* natural language processing
- NN. *See* neural networks
- nodes
  - circle graph, 292
  - concept hierarchy, 20
  - extracted literature as, 190
  - internal, 23, 180
  - radiating, 227
  - relatedness of, 227
  - sibling, 22
  - tree structure of, 182, 195
  - vertices as, 33
- nominals, predicate, 110–111

- noun groups, 107–108
- noun phrases, 97, 113, 114, 115. *See also* generic
  - noun phrases; phrases, coreferring; pronoun resolution engine; proper noun phrases
- definite, 117
- demonstrative, 117
- indefinite, 117
- NPs
  - base, 154
  - chunking of, 154
- Nymble, 149–152
  - experimental evaluation of, 152
  - HMM topology of, 150
  - tokenization and, 150
- object tree. *See* O-Tree
- objects
  - hierarchical structure among, 123
  - structured, 95
- OLDMEDLINE, 12
- one-anaphora, 111
- online categorization, 67
- ontologies, 8, 42, 43
  - commercial, 45
  - creation, 244–248, 275
  - DAG, 197
  - domain, 42–43, 44, 51
  - external, 100
  - Gene Ontology Consortium, 275
  - hierarchical forms generated by, 46
- open-ended architecture, 116
- operations
  - browsing-support, 203
  - link, 203–204
  - preprocessing, 202–204
  - presentation, 204
  - search, 203
- optimization, 6
  - algorithm, 246
  - clustering, 85
  - problems, 84, 139
- orderings, partial, 201–202
- ordinal anaphora, 111
- orthographic features, 100
- O-Tree(s), 123
  - algorithm, 124–125
  - documents structured as, 125
- output concepts, 156
- overabundance
  - pattern, 9, 189
  - problem of, 9, 179
- overlapped matches, 101
- OWL, 275
- pairs
  - tag–tag, 153
  - tag–word, 153
- pajek
  - block modeling of, 268
  - scope of, 271
  - shrinking option of, 258
  - Web site, 271
- Palka system, 166
- paradigm, input–output, 13
- parameter
  - loss ratio, 74
  - maximum likelihood, 91
  - search, 180
- parameterization, 11, 178
- parse tree, 137
- parsing
  - problem, 138
  - shallow, 61, 107–108, 154–155, 167
  - syntactic, 59, 60–61
  - XML, 116
- partial orderings, 201–202
- partitioning, 257–270
- part-of-speech (POS) tagging, 59, 60, 113, 114, 156, 285. *See also* Brown Corpus tag set; lemmas
  - categories of, 60
  - conditional random fields with, 153–154
  - external, 163
  - HMM-based, 156
- part-whole coreference, 112
- patent(s)
  - analysis, 295, 298
  - documents, 304
  - managers, 307
  - search, 274, 295
  - strategy, 295
  - trends in issued, 303–307
- Patent Researcher, 295
  - application usage scenarios, 300
  - architecture/functionality of, 296–300
  - bundled terms, 304
  - constraints supported by, 298–299
  - core text mining operations and, 298–299
  - data for, 297
  - DIAL language and, 297
  - GeneWays comparison with, 307
  - GUI of, 299–300
  - implementation of, 296, 297
  - knowledge discovery support by, 299
  - preprocessing operations, 297–298
  - presentation layer, 299–300
  - queries, 299
  - refinement constraints, 298–299
  - Taxonomy Chooser interface, 297
  - trend analysis capabilities of, 303
  - visualization tools, 299–300, 301
- path(s), 243, 248–249

- pattern(s)
  - browsing, 14
  - collocation, 115
  - concept, 10
  - concept occurrence, 19
  - DIAL language text, 317–318
  - discovery, 5
  - distribution-based, 29, 32, 301
  - Document Explorer search, 236
  - elements, 323–327
  - interesting, 29–30
  - knowledge discovery, 14
  - matching, 270, 322–323
  - network, 242
  - operators, 323
  - overabundance, 9, 189
  - RlogF, 173
  - search for, 8–10
  - sequential, 41
  - text, 317–318, 323
  - text mining, 1, 19
  - textual data, 40
  - unsuspected, 191
  - user's knowledge of, 36
- PCFGs, disambiguation ability of, 156
- PDF files, 3
- percentage thresholds, 38, 39
- perceptual grouping, 59, 123–124
- performance measures, 79
- Perl script, 164
- PersonAffiliation relation, 162–163
- Phillips, W., 171
- phrases, coreferring, 109. *See also* generic noun
  - phrases; noun phrases; proper noun phrases
- PieSky* software, 231
- plan, hyperbolic non-Euclidean, 217
- pleonastic, 110
- PNP. *See* proper noun phrases
- polysemy, 45, 69
- POS tags. *See* part-of-speech tagging
- power centrality, 254–255
- predicate(s)
  - nominals, 110–111
  - unary/binary, 16
- preference
  - collocation pattern, 115
  - domain terminology, 116
  - section heading, 115
- prefix
  - lengthening of, 149
  - splitting of, 149
- preprocessing methodologies, xi, 2, 57
  - algorithms, 57
  - architecture, 58
  - categorizing, 57
  - GeneWays', 308–310
  - Patent Researcher's, 297–298
  - task oriented, 13, 57
  - varieties of, 57
- presentation layer, 185–186
  - components, 14
  - elements of, 1, 14
  - importance of, 10–11
  - Industry Analyzer, 285–288
  - interface, 186
  - Patent Researcher, 299–300
  - text mining system's, 10
  - utilities, 193
- presentation operations, 204
- prestige, types of, 249
- Princeton University, 43
- priors
  - Gaussian, 71
  - Laplace, 71, 72
- probabilistic classifiers, 71, 78
- probabilistic extraction algorithm, 121
- probabilistic generative process, 131
- probabilistic models, 131
- probability
  - conditional, 71, 142, 143
  - context-dependent, 149, 152
  - emission, 132, 150
  - marginal, 71
  - transition, 132, 141, 150, 151
- problem(s)
  - bootstrapping, 172
  - categorization, 82
  - clustering, 84–85
  - CRF, 143
  - data sparseness, 136–137, 148
  - definition, 122–123
  - document sorting, 65
  - HMM's, 132
  - label bias, 142
  - optimization, 84, 139
  - overabundance, 9, 179
  - parsing, 138
  - sets, 194
  - tasks dependent on, 58, 59, 61–62
  - TC, 69
  - text categorization, 69, 79
  - unsolved, 58
- procedure
  - C4.5, 73
  - CART, 73
  - forward-backward, 132–133
  - ID3, 73
- process, probabilistic generative, 131
- processing
  - linguistic, 283
  - random, 131
  - themes related to, 131
- profiles, user, 11
- pronominal anaphora, 110



- pronominal resolution, 112
- pronoun resolution engine, 113
- proper names, 97
  - coreference, 110
  - identification, 106–107
- proper noun phrases (PNP), 171, 172. *See also*
  - generic noun phrases; noun phrases;
  - phrases, coreferring; pronoun resolution engine
- proportional thresholding, 67
- proportions, 19
  - concept, 22, 29
  - interestingness and, 29–30
- protocols
  - RDF, 194
  - XML-oriented, 194
- prototyping, 207
- proximity, 191
- pruning, 73, 178
- PubMed, 2, 11, 275
  - scope of, 2
  - Web site, 12
- pull-down boxes, 276, 278
- quality constraints, 186
- query
  - association-discovery, 45, 46
  - canned, 278
  - choosing entities for, 275
  - clustering and, 83
  - constraints, 278
  - construction of, 278
  - distribution-type, 205, 292
  - engines, 16
  - expressions, 45
  - GUI driven, 284
  - Industry Analyzer event-type, 290
  - interpreters, 10
  - KDTL, 52–54, 55, 236
  - languages, 10, 14, 51–52, 177
  - lists of, 276
  - parameterization of, 11, 178
  - Patent Researcher, 299
  - preset, 274, 276
  - proportion-type, 205
  - result sets and, 45
  - support for, 179
  - tables, 23
  - templates for, 278
  - trend analysis, 304
  - user's, 13
- query languages, 10, 14
  - accessing, 177, 186–187
  - FACT's, 46
  - interfaces, 10
  - parameterization, 178
  - text mining, 51–52
- RDF protocols, 194
- redundancy
  - constraints, 186
  - filters, 201–202
- Reed Elsevier company, 275
- reference, immediate, 115
- referential distance, 116
- refinement
  - constraints, 11, 14, 19–41, 191, 284–285, 298–299
  - techniques, 14–17, 186
- regression methods, 74
- Reingold, E., 231, 246. *See also* FR method
- reiteration, lexical, 115
- relatedness
  - node, 227
  - semantic, 69
- relationship(s)
  - building, 108
  - categorization by, 45
  - context, 32, 33
  - co-occurrence, 12
  - data, 2
  - extraction, 156, 164–166
  - meaningful, content-bearing, 94
  - meronymy/holonymy, 112
  - PersonAffiliation, 162–163
  - rule attributes, 42
  - tagged, 164
  - temporal context, 35
  - term, 275
- relativity, 191
- representations
  - 2-D, 219
  - bag-of-words document, 89
  - binary document, 73
  - character's, 5
  - concept-level, 7
  - document's, 4, 5, 6, 7, 58, 68
  - term-based, 6, 7
  - word-level, 6
- research
  - deviation detection, 32
  - enhancing speed/efficiency of, 2
  - life sciences,
  - text mining and patent, 273
- resolution. *See also* anaphora resolution;
  - coreference, resolution
  - coreference, 109, 112
  - pronominal, 112
- result sets, 45
- retrieval
  - cluster-based, 84
  - document, 179
  - information, 1, 2, 62, 82
- Reuters newswire, 4, 31, 70
- RHS. *See* right-hand side
- right brain stimulation, 191

- right-hand side (RHS), 26, 45, 200
- Riloff, Ellen, 166, 168, 169, 171
- Ripper algorithm, 74, 298
- RlogF pattern, 173
- Rocchio classifier, 74–75
- ROLE
  - relation, 165
  - rules, 165–166
- rule(s)
  - associations, 24, 25, 27, 182
  - averaging, 265
  - constraints, 327–328
  - DNF, 73
  - learners, 74
  - maximal association, 27, 40
  - ROLE, 165–166
  - tagging, 120
  - TEG syntax, 156
- Rzhetsky, A., 310
- salience algorithm, 114
- Sarkar, M., 228
- scaling. *See* generalized iterative scaling
- scanner properties, 320, 327
- scatter/gather browsing method, 83
- scattering, 83
- scenario templates (STs), 99
- SCFG. *See* stochastic context-free grammars
- schemes
  - classification, 131
  - TF-IDF, 68
  - weighting, 68
- search. *See also* Google
  - algorithms, 36, 178, 236
  - association rules, 36
  - brute force, 9
  - constraints, 178, 203
  - DIAL language, 321
  - Document Explorer patterns of, 236
  - engines, 82, 199
  - expanding parameters of, 180
  - fuzzy, 184
  - improving, 82–83
  - Industry Analyzer, 291
  - leveraging knowledge from previous, 36
  - operations, 203
  - parameters, 180
  - patent, 274, 295
  - for patterns, 8–10
  - precision, 83
  - task conditioning, 178
  - for trends, 8–10
- selection. *See* feature, selection
- self-organizing maps (SOMs). *See also* WEBSOM
  - algorithm generation of, 213, 216–217
  - multiple-lattice, 219
- self-loops, 244
- semantic features, 100
- semantic lexicon, 169, 170
- semantic relatedness, 69
- sentences
  - concept, 321
  - linguistic analysis of, 109
- sequential patterns, 30, 41
- sets. *See also* concept sets
  - answer, 1, 23
  - association rules involving, 200
  - Brown Corpus tag, 60
  - frequent and near frequent, 19, 25
  - frequent concept, 9, 23–24
  - identical, 111
  - POS tag, 60
  - result, 45
  - $\sigma$ -cover, 24
  - test, 67, 100
  - test document, 79
  - training, 68, 79, 100, 118
  - validation, 68, 75
- shallow parsing, 61, 107–108, 154–155, 167
- shrinkage, 136
  - defined, 136
  - hierarchies, 148
  - ngram, 163
  - technique, 148, 161
- shuffling algorithms, 85
- sibling node, 22
- similarity
  - cosine, 90, 200, 201
  - function, 84, 200–201
  - measures, 85
- simple concept association graphs, 200–201
- simple concept graphs, 195–205, 239, 286, 294
- simulated annealing, 247
- single-label categorization, 67
- singleton
  - $\sigma$ -covers, 24
  - vertex, 198
- singular value decomposition (SVD), 89–90, 91
- smoothing, 136
- social networks, 242
- soft mark-up language, 3
- software
  - browsing, 177
  - corporate intelligence, 273
  - Insight, 217
  - libraries, 207
  - link analysis, 271–272
  - NetMiner, 272
  - PieSky, 231
  - protein interaction analysis, 273
  - search engine, 199
  - StarTree Studio, 217
- SOMs. *See* self-organizing maps
- Soon's string match feature (SOON STR), 118
- sparse network, 244
- sparseness, 72
  - data, 136–137, 148
  - training data, 136–137

- sparsity. *See* feature, sparsity
- spring embedding, 231 *See also*
  - networks
  - algorithms, 245
  - network graphs, 231
- StarTree Studio* software, 217
- states
  - background, 149
  - HMM's classes of, 147
- stimulation, right brain, 191
- stochastic context-free grammars (SCFG), 131, 137
  - defined, 138
  - information extraction and, 155–166
  - using, 137–138
- stop words, 68
- strategy
  - divide-and-conquer, 58
  - patent, 295
- string, 138
  - constants, 324
- strong components, 260
- structural equivalence, 261
- structural mapping, 125–127
- structured objects, 95
- STs. *See* scenario templates
- sublanguages, 138
- subtasks, 123–124
- suffix
  - lengthening of, 149
  - splitting of, 149
- sum-of-squares, 29
- supervised learning, 70
- support, 24, 25, 249
  - query, 179
  - thresholds, 181
  - vectors, 76
- support vector machines (SVM), 76–77, 78 76–77, 78
- SVD. *See* singular value decomposition
- SVM. *See* support vector machines
- Swiss-Prot database, 308
- symbolic classifiers, 72
- symbols, terminal, 156
- SYNDICATE system, 18
- Synonym/Homonym resolver, 309
- synonymy, 45, 69
- syntactic analysis, 105
- syntactic heuristics, 171, 172
- syntactic parsing, 59, 60–61
- syntactical constraints, 186
- syntax, TEG rulebook, 156
- system(s). *See also* AutoSlog-TS system;
  - CONSTRUE system; Explora system;
  - GENomics Information Extraction System;
  - hybrid system; MedLEE medical NLP system;
  - Palka system; SYNDICATE system;
  - text mining systems; TEXTRISE system
- architects, 17
- architecture, 46–47, 186
- thresholds defined by, 229
- table(s)
  - of contents, 83
  - joins, 1
  - query, 23
- tagging. *See also* part-of-speech tagging
  - chunk, 155
  - documents, 94
  - MUC style, 100
  - POS, 283
  - rules, 120
- tag–tag pairs, 153
- tag–word pairs, 153
- target string
  - lengthening of, 149
  - splitting of, 149
- task(s). *See also* coreference task; subtasks;
  - template element tasks; template relationship task; visual information extraction task
  - AI, 64
  - algorithms, 58
  - clustering, 82–84
  - documents structured by, 57
  - entity extraction, 150, 156
  - NE, 96
  - preprocessing by, 13, 57
  - problem dependent, 58, 59, 61–62
  - search, 178
  - text categorization, 66
  - TR, 99
- taxonomies, 8, 42, 180
  - classic, 185
  - concept, 195
  - editors, 183–184
  - maintaining, 183
  - roles of, 182
- Taxonomy Chooser interface, 297
- TC. *See* text categorization
- TEG. *See* trainable extraction grammar
- template element (TE) tasks, 98
- template relationship (TR) task, 99
- templates, 123, 127–128, 278
- temporal context graphs, 30, 32, 35
- temporal context relationships, 32, 35
- temporal selection, 35
- term(s), 5–6, 8
  - candidate, 6
  - clustering, 69
  - extraction, 6, 12, 95, 283
  - hierarchy editor, 237–238
  - lemmatized, 6
  - Patent Researcher's bundled, 304
  - relationships, 275
  - tokenized, 6
- terminal symbols, 156

- term-level representations, 7
- termlists, 156
- TEs. *See* template element tasks
- test sets, 67, 100
- text(s)
  - classifiers, 76, 79–80
  - clustering, xi, 89, 91–92
  - comprehension, 95
  - elements extracted from, 96
  - extraction, 96
  - filtering, 65–66
  - fragments, 109
  - natural language, 1
  - pattern, 317, 318, 323
  - tokenization, 320
- text analysis, 146–152
  - clustering tasks in, 82–84
  - CRF's application to, 153–155
- text categorization (TC), 58, 61–62, 64
  - applications, 64, 65–66
  - approaches to, 64
  - automated, 64
  - experiments, 79
  - knowledge engineering approach to, 70
  - machine learning, 70–78
  - NN and, 75
  - problem, 69, 79
  - stop words removed from, 68
  - task, 66
- text mining. *See also* preprocessing methodologies
  - algorithms, 5, 8
  - analysis tools for, 1
  - applications, xi, 8
  - background knowledge and, 8
  - biological pathways,
  - corporate finance and, 273
  - data mining, v, 1, 11
  - defined, x, 1, 13
  - essential task of, 4
  - goals of, 5
  - GUIs for, 177
  - human-centric, 189
  - IE and, 11
  - input–output paradigm for, 13
  - inspiration/direction of, 1
  - introductions to, 11
  - KDD applications, 13
  - life sciences and, 273
  - methodologies, 9
  - patent research and, 273
  - pattern overabundance limitation, 9
  - pattern-discovery algorithms, 1, 5
  - patterns, 19
  - preprocessing operations, 1, 2, 4, 7, 8, 13–14, 57
  - presentation layer elements, 1, 14
  - query languages for, 51–52
  - techniques exploited by, 1
  - visualization tools, 1, 194
- text mining applications, 8
  - corporate finance-oriented, 284
  - Document Explorer, 18
  - Explora system, 18
  - FACT, 18
  - GUIs of, 177
  - horizontal, 307
  - KDT, 18
  - LINDI project, 18
  - SYNDICATE system, 18
  - TEXTRISE system, 18
- text mining systems. *See also* core text mining operations
  - abstract level of, 13
  - architecture of, 13–18
  - background knowledge and, 8, 16, 42, 44
  - baseline distribution for, 22
  - concept proportions and, 29
  - content based browsing with, 10
  - customized profiles with, 11
  - designers of, 221, 275
  - distributions and, 29
  - domain specific data sources, 16
  - early, 30
  - empowering users of, 10
  - front-ends of, 10, 11
  - graphical elements of, 11
  - hypothetical, 19
  - incremental update schemes for, 38
  - practical approach of, 30
  - presentation layer of, 10
  - query engines of, 16
  - refinement constraints, 11
  - refinement techniques, 14–17
  - state-of-the-art, 10, 194
- TEXTRISE system, 18
- textual data, 88–92, 189, 195
- TF-IDF schemes, 68
- thematic hierarchical thesaurus, 65
- themes, processing, 131
- thesaurus
  - MESH, 65
  - names, 325
  - NASA aerospace, 65
  - thematic hierarchical, 65
- three dimensional (3-D) effects, 219–221
- See also* representations, 2-D
  - algorithms, 219
  - challenges of, 220
  - disadvantages of, 221
  - impact of, 221
  - opportunities offered by, 220
- thresholding
  - fixed, 67
  - proportional, 67
- thresholds
  - confidence, 181
  - data, 39

- minconf, 26, 40
- minsup, 26, 40
- percentage, 38, 39
- support, 181
- system-defined, 229
- user-defined, 229
- time-based analysis, 30
- Tipster, 96–101
- Title Browser, 301
- token(s)
  - elements, 327
  - features of, 150, 161
  - ngram generation of, 159, 161
  - .UNK., 152
  - unknown, 152, 161
- tokenization, 59, 60, 104, 106, 107
  - DIAL language text, 320
  - linguistic processing and, 283
  - Nymble and, 150
- tokenizer, external, 161
- tools
  - analysis, 1
  - browsing, 181
  - clustering, 11, 184–185
  - Document Explorer visualization, 236
  - editing, 184
  - graphical, 189
  - hybrid, 221–224
  - hyperbolic tree, 217
  - line graphs as prototyping, 207
  - prototyping, 207
  - visualization, 1, 10, 14, 192, 194, 227–228, 294
    - 1, 10, 14, 192, 194, 226–227, 292
- TR. *See* template relationship task
- trainable extraction grammar (TEG), 155, 156
  - accuracy of, 165
  - experimental evaluation of, 164
  - extractor, 164
  - grammar, 157, 158
  - as hybrid system, 156
  - rulebook syntax, 156
  - training, 158–161
- training
  - classifiers, 79
  - CRF's, 144
  - examples, 117
  - HMM, 135–136
  - MEMM, 141
  - sets, 68, 79, 100, 118
  - TEG, 158–161
- transmission
  - emission cycle, 131
  - matrix, 143
- tree(s). *See also* minimal spanning tree
  - binary, 73
  - browsing, 15
  - hierarchical, 42, 195
  - hyperbolic, 217–219
  - node structure of, 182
  - parse, 137
  - pruning, 73, 178
- trend(s)
  - analysis, 9, 30–31, 41, 299, 303
  - graphs, 30, 32, 35, 239
  - patent, 303–307
  - search for, 8–10
- trigger-constraint functions, 153
- trigrams, 5
- tuple dimension, 5
- two-mode network, 244
- UCINET, 271–272
- UMLS Metathaurus. *See* Unified Medical Language System Metathesaurus
- unary predicates, 16
- undirected networks, 260
- Unified Medical Language System (UMLS) Metathesaurus, 282
- uniformity, 152
- United States Patent and Trademark Office, 297, 303
- unknown tokens, 152, 161
- .UNK. tokens, 152
- user(s)
  - browsing by, 10, 13
  - clustering guided by, 83
  - customizing profiles of, 11
  - empowering, 10
  - groups, 194
  - interactivity of, 179, 189
  - M-support and, 28
  - pattern knowledge of, 36
  - querying by, 13
  - thresholds defined by, 229
  - values identified by, 26
- user-identified values
  - minconf*, 26
  - minsup*, 26
- utilities, 193
- validation sets, 68, 75
- variable
  - backward, 133
  - forward, 132, 141
- vector(s)
  - feature, 68
  - formats, 7
  - global feature, 142
  - original document, 90
  - space model, 85
  - support, 76
  - weight, 142
- verbs, indicating, 115
- vertices, 33, 258, 264
- VIE task. *See* visual information extraction task

- visual information extraction (VIE) task, 122, 123, 128
- visual techniques, 194–225
- visualization. *See also* circle graphs
  - 3-D, 219
  - approaches, 189, 191, 279
  - assigning colors to, 279
  - capabilities, 274
  - circle graph, 292
  - DAG techniques of, 198
  - data, 217
  - Document Explorer tools for, 236, 238
  - hyperbolic tree, 217
  - interface, 191
  - link analysis and, 225
  - Patent Researcher's tools for, 299–300, 301
  - specialized approach to, 225
  - tools, 1, 10, 14, 192, 194, 226–227, 292
  - user interactivity and, 189
- Viterbi algorithm, 133–134, 138, 141
- vocabulary, controlled, 65, 275
- walk, 243
- Washington Post* Web site, 244
- weak components, 260
- weak hypothesis, 77
- weak learner, 77
- Web pages
  - hierarchical categorization of, 66
  - HTML and, 3
  - hypertextual nature of, 66
- Web site(s)
  - FBI, 244
  - Kevin Bacon game, 248
  - pajek, 271
  - PubMed, 12
  - UCINET, 271–272
  - U.S. Patent and Trademark Office, 297, 303
  - Washington Post, 244
- WEBSOM, 213–215. *See also* self-organizing maps
  - advantages of, 215
  - zoomable interface of, 215
- weight vector, 142
- weighted linear combination, 77
- weights
  - binary, 68
  - giving, 68
- WHISK algorithm, 119
- word stems, 4
- wordclass names, 324–325
- word-level representations, 6
- WordNet, 43, 44, 50, 51, 112
- word-processing files, 3
- words, 5–6, 8
  - identifying single, 6
  - POS tag categorization of, 60
  - scanning, 106
  - stop, 68
  - syntactic role of, 58
  - synthetic features v. naturally occurring, 69
- workbench, evaluation, 116
- WYSIWYG HTML editor, 3
- Xerox PARC, 217
- XML
  - parsing, 116
  - protocol, 194
- Yang, Y., 76
- Zhou, M., 197, 198
- zoning module. *See* tokenization
- zoomability, 191