

Index

Symbols

2×2 contingency matrix 747

A

abnormal detection 700, 707
 AdaBoost 111, 115, 116, 119, 122, 123, 124, 125, 126
 anaphora resolution 44
 anomalies detection 412
 ant colony optimization 169, 171, 174, 179
 application tuning 413
 association rule mining 306, 313, 368, 715
 association rules 98, 242, 244, 246, 262, 267, 312, 313, 357, 363, 368
 Atype 585, 594, 595
 Australian Department of Health and Ageing (DoHA) 786
 authorship characterization 707
 authorship identification 707
 automated text categorization 331
 automatic annotation 322
 automatic term recognition 529

B

back-propagation neural networks (BPNNs) 201–218
 agents 214–215
 Web text mining system 203–212
 feature vector conversion 209
 framework 203–204
 learning mechanism 210–211
 limitations 211–212
 main processes 204–208
 bag-of-words representation 1
 bag of words 789

Bayesian 574, 575, 576, 577, 578, 582, 583, 601, 603
 learning 421
 BayesQA 574, 575, 582, 599, 603
 BayesWN 574, 577, 583, 603
 BioLiterature 323, 328, 329
 biomedical databases 329
 biomedical research 314
 BioOntologies 314–330
 BioOntology 316, 322, 323, 326, 327, 329
 blocking-based techniques 474
 blogosphere 646, 647, 651, 656, 657, 660, 662, 663, 665, 668
 boosting 111, 112, 113, 115, 118, 119, 122, 123, 124, 125, 126, 127, 133, 187, 338, 601
 BoW approach 4
 bucket 274, 285, 287
 bucket-based histogram 287

C

CASE tool 402, 404, 406
 classification, naïve Bayes 788
 classification error 109
 classification rule mining 110
 classifications, multiple 798
 classification uncertainty 127
 clause cube 299
 clinical document architecture (CDA) 683
 clone detection 627, 644
 Cluster category 801
 clustering 424
 crossed clustering approach 430
 methods 429
 clustering of a dataset 188
 cluster prototypes 723

- clusters 801
 - co-clustering 723
 - code compaction 644
 - communication-channel congestion control 412
 - compare suite 766, 769, 770, 771, 784
 - complex terms 529
 - concept-based text mining x, xxii, 346, 358
 - conceptive landmark 469
 - Conceptual logs 407
 - conceptual logs 408
 - consensus function 188
 - consumer informatics xiii, xxx, 758, 764
 - content-based image retrieval (CBIR) 110
 - content metadata 36
 - Content mining 605
 - content personalization 402
 - content similarity 374, 375, 385, 746
 - content similarity measure 746
 - contents in XML document 247, 271
 - content unit 414
 - control data flow graph (CDFG) 644
 - course syllabus 61
 - crossed dynamic algorithm 430
- D**
- data cube ix, xx, 288, 299
 - data mining, incremental or interactive 448
 - data mining techniques 249, 250, 252, 260, 266, 267
 - data modification 473
 - data partitioning 472
 - data pre-processing techniques 273
 - data restriction techniques 474
 - data stream pre-processing 287
 - data warehousing 245, 285, 299, 367, 467, 479, 644, 692, 722
 - deployment landmark 469
 - descriptor 344
 - designing Web applications x, xxiii, 401
 - dicing ix, xx, 288, 294, 299
 - dimensionality reduction 567, 568, 569, 711
 - document vii, viii, xv, xviii, 2, 6, 17, 18, 20, 21, 22, 24, 35, 178, 181, 182, 35, 36, 93, 146, 166, 167, 172, 177, 178, 180, 179, 180, 23, 180, 187, 188, 1, 23, 165, 181, 187, 188, 204, 210, 231, 245, 247, 251, 267, 268, 271, 272, 304, 306, 327, 344, 485, 487, 491, 492, 493, 503, 508, 544, 549, 550, 560, 569, 610, 675, 683, 706, 743, 744, 745, 746, 754, 820
 - document clustering 181, 182, 187
 - document indexing 2, 344
 - document keyphrases 23–36, 32
 - document metadata 36
 - document representation 3, 13
 - DoHA (Australian Department of Health and Ageing) 786
 - domain-specific keyphrase extraction 36
- E**
- e-health xii, xxviii, 670, 683
 - EBIMed 324
 - electronic commerce 449
 - EM algorithm 80, 92, 114, 117, 125, 127
 - embedded system 644
 - EMOO 47
 - ensemble scheme 181, 184
 - entity-relationship [E/R] 404
 - EUROVOC 334
 - evolutionary multi-objective optimisation (EMOO) 47
 - extensible markup language (XML) 205
 - external measures 747
 - extrinsic evaluation 747
- F**
- feature selection 19, 21, 22, 74, 206, 305, 692
 - fingerprinting 636, 644
 - flocking model 180
 - focus word 603
 - folksonomy 668
 - frequent pattern mining 227, 228, 229, 235, 239, 241, 243
 - frequent patterns 247, 271, 272
 - frequent patterns mining 247, 271, 272
 - frequent word sequence 5
 - full syllabus 74

Index

G

Gaussian kernel 794
Gene 317, 319, 321, 323, 324, 325, 328, 330, 691, 692, 693, 751, 752, 754, 756, 757, 833
gene expression 319, 330, 691, 692, 751, 752, 754, 757
gene ontology 317
General Purpose Techniques (GPT) 475
genetic algorithms (GA) 38, 40
genome annotation 752, 757
gold standard 747
GoPubMed 323

H

Health Insurance Commission (HIC) 786
health leven seven (HL7) 683
HIC (Health Insurance Commission) 786
HIC category 801
hierarchical clustering 180
hierarchical structure 273, 274
histogram 107, 286, 287
HITS Algorithm 617
homogeneity 612
hybrid scheme 181, 185
hypergraph 412
hypertext makeup language (HTML) 205
hypertexts 404

I

IE 38, 39
Imbalance in Data 723
incremental Web traversal pattern mining (Incremental WTP) 455
index 325, 336, 544, 747
inflection 545
information extraction (IE) 38
information filtering system 499
information retrieval 19, 20, 21, 32, 36, 76, 93, 109, 126, 138, 147, 164, 180, 187, 199, 270, 303, 312, 313, 322, 329, 342, 344, 347, 349, 357, 384, 420, 498, 499, 501, 543, 544, 545, 547, 559, 560, 562, 567, 568, 569, 572, 573, 600, 602, 605, 622, 676, 681,

705, 721, 735, 744, 745, 749, 750, 757, 764, 780, 781, 783, 805, 838
information retrieval (IR) 38, 39, 545, 749, 757
inlink and outlink 385
interactive Web traversal pattern mining (Incremental WTP) 459
interestingness 48, 312, 313
internal measures 747
intrinsic evaluation 747

K

KDD process 408
KDD scenario 408
KDD scenarios 402, 409
KDT 38, 39
kernel, Gaussian 794
kernel, latent semantic 795
kernel, power 794
keyphrase assignment 24
keyphrase extraction 24
keyphrase identification program (KIP) 23
Kintsch's Predication 48
knowledge
 discovery 418
knowledge base assisted incremental sequential pattern (KISP) 453
Knowledge Discovery 273, 275, 285
knowledge discovery 249, 250, 264, 266, 271, 604
knowledge discovery from databases (KDD) 38
knowledge discovery from texts (KDT) 38, 39

L

latent semantic analysis (LSA) 43
Latent Semantic Kernel 795
latent semantic kernel (LSK) 787
latent semantic space 569
lemmatization 534, 535, 541, 544, 545
lexico-syntactic patterns 422
linear programming problem 683
linguistic data 288, 289, 291, 292, 295, 298, 299
link similarity 374, 376, 385
logical observation identifiers names and codes

(LOINC) 683
 lymphoblastic leukemia 689, 693

M

maximal Frequent word sequence (MFS) 3
 medical literature analysis and retrieval system
 online (MEDLINE) 764
 medical subject headings (MeSH) 764
 Medicare Benefit Schedule (MBS) 786
 Megaputer TextAnalyst 773, 784
 MGED ontology 319
 micro-array 693
 minimal spanning tree 400
 minimum description length
 684, 685, 686, 688, 693
 mining Web applications x, xxiii, 401
 mining XML 227, 228, 235, 243
 model-based methods 715, 723
 model testing 74
 model training 74
 molecular biology 138, 330, 838
 MRR 580, 581, 582, 598, 599
 multi-agent-based Web text mining system
 212–214
 implementation 214
 structure 212–213
 multi-resolution analysis 110
 multiple classifications 798
 multitarget classification 693

N

naïve bayes viii, xvii, 2, 13, 111, 112, 11
 3, 114, 115, 111, 115, 114, 115, 11
 6, 117, 118, 119, 120, 121, 122, 1
 23, 124, 125, 126, 127, 137, 339, 68
 6, 687, 788
 naïve Bayes (NB) algorithm 202
 Naïve Bayes classification 788
 NAL Agricultural Thesaurus 335
 natural-language processing 43
 natural language processing 19, 20, 35, 312,
 322, 344, 527, 528, 602, 745, 749, 7
 50, 757, 784
 natural language processing (NLP)
 204, 322, 749, 757
 Naïve Bayes 66

Newsgroup mining structure 615
 nio-informatics 693
 noise addition techniques 473
 noun phrase 26
 noun phrases 23
 noun phrases extractor 27
 novelty 40, 301, 306, 308, 310, 311, 312,
 313
 NsySQLQA 574, 575, 582, 583, 596, 599,
 603
 nuggets 56

O

Off-Line Phase 814
 Online Phase 819
 ontologies 403
 ontology x, xxiii, 163, 199, 244, 316, 317,
 318, 319, 320, 321, 322, 323, 324,
 325, 327, 328, 329, 330, 402, 418, 4
 21, 422, 423, 425, 433, 435, 440, 44
 1, 443, 444, 445, 499, 528, 752, 756
 , 764, 833
 building
 and Web site description 435
 construction 421
 method 425
 evolution 422
 management 418–447
 Web usage mining 423
 open biomedical ontologies 321

P

page annotation 402, 404
 Pareto dominance 50
 part-of-speech (POS) 43
 part-of-speech tagger 27
 partial syllabus 74
 particle swarm optimization
 166, 169, 170, 179, 180
 partitioning clustering 180
 partition of a dataset 188
 path traversal pattern mining 451
 pattern 410
 pattern detection 110
 pattern discovery 615
 pattern recognition 110, 125, 177, 188, 746

Index

- patterns 38, 402, 409, 415
- patterns statistically 410
- people search 385
- performance evaluation measure 747
- Personalization 604
- personalization 92, 313, 365, 366, 367, 399, 420, 445, 604, 619, 655
- person search 385
- positive set 94
- power kernel 794
- pre-processing 403
- precision 32
- predictive model 139
- principal component analysis 556, 568, 685, 693
- principal components analysis (PCA) 552, 570
- privacy violation 470
- probabilistic latent semantic analysis (PLSA) 403
- probability based term weighting 13
- procedural abstraction 629, 643, 644
- prospective landmark 470
- protein classification 139, 140
- prototype 188, 370, 383, 446
- PU learning 94
- Q**
- query expansion 356, 545
- query term/keyword 545
- R**
- range query 287
- rank correlation coefficient 747
- recall 32
- regularity extraction 631, 644
- reliable multicast transport protocol (RMTP) 220
- Reuters-21578 12, 13, 14, 15, 16, 120, 200, 568
- Roc-SVM 75, 78, 85, 87, 92, 94
- rotation 292, 293, 299
- S**
- S-EM 75, 78, 80, 81, 82, 83, 85, 87, 90, 92, 94
- sanitization-based techniques 474
- SAS text miner 202, 222, 766, 769, 770, 771, 773, 780, 784
- secure multi-party computation (SMC) 472
- selective sampling 115, 119, 127
- selector 603
- self-organizing map (SOM) 203
- self organizing maps of kohonen 200
- semantic expansion 358
- semantic heterogeneity 723
- semantic measure 164
- semantic similarity 47
- Semantic Web 442
 - mining 442
 - visualisation 442
- SemSim 44
- sequential pattern mining 424, 452
- similarity transformation 110
- singular vector decomposition (SVD) 44
- slicing ix, xx, 288, 291, 293, 294, 299
- sliding window 282, 287
- social “friendship” network 668
- social network analysis 668, 697, 698, 707
- space transformation techniques 473
- SPEA 50
- SPSS mining for clementine 784
- statistical language models 25
- stemming 385, 540, 545
- stemmisation technique 789
- stopwords 535, 545
- strength Pareto evolutionary algorithm (SPEA) 50
- structure mining 605
- structures in XML document 247
- subgraph 234, 247, 271, 644
- subtree 236, 247, 268, 272
- suffix tree 645
- supervised classification 93, 344
- supervised learning 64
- support vector machin 788
- support vector machine methodology 792
- support vector machines (SVM) 61–74
- swarm intelligence 165, 169, 178, 180
- syllabi 61
- syllabi, defined classes 62
- syllabus component 74

syllabus entry page 74
 synonymy 44, 256, 723
 systematized nomenclature of medicine
 (SNOMED) 683
 systems biology 757

T

template 408, 409
 temporal analysis 353, 358
 termhood xi, xxv, 500, 502, 507, 521, 522,
 526, 529
 termhood evidences 529
 term weight 385
 term weighting 12
 term weighting schemes 10
 text categorization viii, xvii, 18, 19, 21, 22,
 128, 140, 199, 336, 337, 342, 343,
 344, 445, 602, 745, 784
 text classification 1–22
 text classifiers 331
 text coherence 49
 text mining (TM) 39
 thesauri, definition of 334
 thesaurus 332
 thesaurus, indexing 337
 thesaurus, text categorization 336
 thesaurus-based automatic indexing 331–345
 thesaurus formalization 333
 three-dimensional array
 291, 292, 293, 294, 299
 TM 39
 tokenizer 27
 tourism 433
 traffic analyzers 403
 traversal sequence 449
 two-step strategy of PU learning 94

U

Unified Medical Language System 320
 uniform resource identifier (URI) 164
 unithood xi, xxv, 500, 501, 502, 506, 514,
 521, 526, 529
 unlabeled set 95
 unsupervised classification 344
 usage-mining preprocessing phase 408
 Usage mining 605

Usenet Site 607
 user
 profile 420
 session 421, 427
 user modeling 446, 622, 683
 user profile 683
 user profiling x, xxii, 359, 360, 368

V

vector space methods 570
 vector space model 147, 166, 179, 180, 304,
 340, 344, 569
 vector space model (VSM) 202
 visual text 766, 769, 770, 780, 784
 vocabulary mapping 765
 vocabulary problem 358

W

wavelets 287
 Web 408
 -based
 information
 systems 418
 content mining 420
 log 426, 441
 preprocessing 435
 mining 420, 423, 442
 structure
 mining 420
 text mining 201–218
 usage 441
 mining x, xxiii, 418, 420, 441
 Web-log 402, 403
 Web application development 404
 Web content analysis 707
 Web Data Techniques (WDT) 475
 Web graph 400
 Web image search engine 110
 Web link analysis 707
 Web logs 448
 Web mining 448–467
 WebML 406
 WebML method 404
 Web modeling language (WebML) 402
 Web site 819
 Web structure mining 360, 400
 Web traversal x–xiv, xxiv–xxxii, 448–467

Index

Web traversal patterns, mining of 453
Web usage-mining 403
Web usage mining x, xxii, xxiii, 270, 359, 360, 363, 361, 359, 364, 363, 364, 365, 366, 368, 400, 402, 418, 428
WordNet 40
Wordnet 194, 199, 200, 349, 528, 601
WordStat 766, 769, 770, 777, 778, 784
word stemming 385

X

XML 407
XML document handling 249, 253
XML frequent content mining 231, 247, 272
XML frequent patterns mining 247, 272
XML frequent structures mining 247, 272
XML standardization 227