Best trail algorithm, 245–252
convergence stage, 248
effective view navigation, 245–246
exploration stage, 248
nav-search, 249
probabilistic best first search algorithm, 248
trail engine, developing, 246–252
trail search, 249
visual search, 251
web usage mining for personalization, 246
Betweenness centrality, 322–323
Bias and demand, trade-off between, 160–161
Bibliometrics, 127
BigTable database system, 87
Binary relevance assessment, 94
Bing (www.bing.com), 8, 70–72
 Bipartite graph, 121, 325
BitTorrent File Distribution, 331–332
BitTorrent protocol, 329
BlackBerry (www.blackberry.com), 283
Blending, 343
Blind feedback, 183
Block, 324
Blockmodeling, 324
BlockRank, 185
Blogs/Weblogs, 124, 347–352
Aardvark (www.vark.com), 352
blogrolling, 348
blogspace, 348–349
ChaCha (www.chacha.com), 352
for machine learning algorithms testing, 349
microblogging, 350–352
real-timeweb, 350–352
spreading ideas via, 349–350
Technorati (www.technorati.com), 348
Blogspace, 348–349
Bookmarking, 379–389
Delicious (http://delicious.com) for, 382
Bookmarks tool, 25, 216–219
bookmarklet, 218
clustering algorithm, 217
k means clustering algorithm, 217
organizing into category, 218
Borda count, 170
Bounce rate, 159
Breadth-first crawl, 83
Breadth-first search algorithm, 248
Breadth-first search, 249
Breadcrumb navigation, 221–222
breadcrumb trail, 221
location breadcrumbs, 221
path breadcrumbs, 221
Browsers, mobile web, 282–283
Browsing, 35, 385–388
basic browser tools, 213–214
frustration in, 211–213
404 error—web page not found, 212
HTML, 211
hyperlinks, 211–212
pop-up ads, 212
referential integrity problem, 212
surfing, 211–212
web site design and usability, 211–213
Bursty, 349
Buttons, 224
C
Caching queries, 102
Candidate set, 300
CAPTCHA, 199–200
Capture-recapture method, 12
Card, 224
Carrot (www.carrot-search.com), 258, 261
Cascading style sheets (CSS), 394
Categorization of web content, 150–151
CDNow (www.cdnow.com), 347
Centralization, 226–227, 322–324
betweenness centrality, 322
closeness centrality, 322
degree centrality, 322
peripheral, 323
star, 323
Centralized P2P networks, 327–328
ChaCha (www.chacha.com), 352
Cinematch, 342
Citation analysis, 127–129
CiteULike (www.citeulike.org), 383
Classifying search results, 175–178
Classifying tags, 389
Click-based approach, 179
Clickbots, 167
Click-distance, 291–292
Click fraud, 166–168
advertiser competitor clicking, 166
automated click fraud, 167
detection, 167
publisher click inflation, 166
Click here, link text, 213
Clickthrough rates (CTRs), 130, 153, 158
Cloaking, 97
Closeness centrality, 322–323
Cloud computing, 398
Clustering, 132, 324, 389
clustering coefficient, 321
Clustering search results, 173–175
goals of, 174
alleviating information overlook, 175
fast topic/subtopic retrieval, 174
topic exploration, 175
Clusty (www.clusty.com), 259
Co-citation network, 128
Collaboration graphs, 313–314
Erdős, 313
Collaborative filtering (CF), 333–347
Amazon.com (www.amazon.com), 333–334
CDNow (www.cdnnow.com), 347
content-based recommendation systems, 339–340
evaluation of, 340–341
item-based, 337
model-based, 338–339
MovieLens (http://movielens.umn.edu), 346
PocketLens, 346
Ringo, 347
scalability of, 341
traditional nearest-neighbor approach to, 344
user-based, 335–337
Collaborative search, 372
Collective intelligence, 399–401
algorithms for, 401–402
community question answering (CQA), 399
crowdsourcing, 400
Collective search, 372
Commerce metrics, 233
Communities within content sharing sites, 382–383
Community search, 372
Compact hypertext markup language (cHTML), 274
Compactness, 228
Comparative sentence, 393
Computer backgammon, 100–101
Computer Chess Programming, 62–65
Concept dictionary, 108
Conditional random fields (CRFs), 245
Condorcet count, 170
Connected triple, 362
Connectivity notion, 226
Content-based image search, 196–198
semantic gap, 196
Content-based recommendation systems, 339–340
Content-based retrieval, 194
Content metrics, 233
Content mining, 230
Content relevance, 94–108
binary relevance assessment, 94
caching queries, 102
date last updated, 104
home page detection, 107
HTML structure weighting, 104–105
inverse document frequency, 99–100
keyword TF–IDF values, computing, 100–102
known item searches, 107
link text, 103–104
nonbinary relevance assessment, 94
non-English queries, 106
phrase matching, 102
query, interpreting, 96
related searches and query suggestions, 107–108
spell checking, 105–106
synonyms, 102–103
term frequency, 96–99, See also individual entry
URL analysis, 104
web pages, processing, 94–96
Content sharing sites, communities within, 382–383
Context-aware proxy-based system (CAPS), 370
Context dimension of reality analytics, 266
Contextual targeting, 157
Continuous Markov chain, 135
Conversion rate (CR), 164, 234
Cookies, 55, 67
flash cookie, 55
Cost per click (CPC), 154
Cost per impression (CPM), 153
Cost per order, 234
Cost per visit, 234
Coverage issue, 11
Craigslist (www.craigslist.org), 397
Crawling the web, 81–85
algorithms, 82–83
breadth-first crawl, 83
First-In-First-Out (FIFO), 83
focused crawlers, 83
Google dance period, 81
refreshing web pages, 84
robots exclusion protocol, 84–85
Cross validation, 242
Crumbs, 223
Cuil (www.cuil.com), 13, 72
Customization, personalization versus, 180
Cybercrime, 422
Cyberspace, mapping, 262

D
Daily behavior vectors, 265
Dangling, 112
Data feeding, 244
Data mining, web, 230–245
perspectives on, 230–231
content mining, 230
structure mining, 231
usage mining, 231
success of a web site, measuring,
231–232
batting average, 232
conversion rate, 232
traffic, 231
visitors number calculating, 232
Data sharing graph, 365
Date last updated, 104
Daum (www.daum.net), 16
Decentralized algorithm, 375
Decentralized P2P Networks, 328–330
Deep web site, 10–11
Degree centrality, 322
Degree distribution, 321
Delicious (http://delicious.com), 382
Derived trails, 47
Design, web site, 211–213
Desktop search, 421
Destination URLs, 79
Digital Bibliography and Library Project (DBLP), 202
Diigo, 218
Direct Hit’s popularity metric, 130–132
Direct navigation, 24–25
bookmarks use, 25
RealNames dot-com, 24
Directories of web content, 150–151
Discounted cumulative gain (DCG), 137
Distributed hash tables (DHTs), 331
Distributions, 352–368, See also
Power-law distributions in web
normal distribution, 352
Document space modification, 132
Dodgeball, 319
Dogpile (www.dogpile.com), 170
Domain name system (DNS), 236
Dominant-SE, 65
Doorway pages, 97
Duplicate pages, 97
Dynamic Markov modeling, 241

E
eBay, 407–412
auction fraud, 411
AuctionWeb, 407
first-digit law, 412
hard close, 407
power buyers, 412
power sellers, 412
reserve price, 409
shill account, 410
Skype, 407
soft close, 407
e-Commerce searching, 159
Eigenbehaviors, 265
Emergent trails, 47
e-Metrics, 233–234
cost per order, 234
cost per visit, 234
repeat order rate, 234
repeat visitor rate, 234
take rate, 234
Ensemble (www.the-ensemble.com), 343
Enterprise search engines, 32
Entity structure, 244
Erdös collaboration graph, 313
404 Error—web page not found, 212
eTesting Labs, 138
e-Tickets, 277
Evaluating search engines, 136–143
   automating web search, 142
   awards, 136
discounted cumulative gain (DCG), 137
cTesting Labs, 138
evaluation metrics, 136–138
   eye tracking studies, 139–141
F-score, 137
inferring ranking algorithms, 142–143
mean reciprocal rank (MRR), 137
normalized discounted cumulative gain (NDCG), 138
Page Hunt, 139
performance measures, 138–139
test collections, 141–142
VeriTest (www.veritest.com), 138
Explicit feedback, 278
Explicit rating, 335
Explicit Semantic Analysis (ESA), 384
Extensible hypertext markup language (XHTML), 274
Extraction, information, See Information extraction
Eye tracking studies, 139–141
F
Facebook (www.facebook.com), 318
Factual queries, 190–192
False positives, 402
Fast topic/subtopic retrieval, 174
FaThum, 297
Fault-tolerant software, 85
Favorites list, See Bookmarks tool
Feature-based opinion mining, 391–392
Fennec (mobile Firefox, https://wiki.mozilla.org/Fennec), 283
FindLaw (www.findlaw.com), 202
Fine-grained heuristic, 242
Firefox, 56
First-digit law, 412
First-In-First-Out (FIFO), 83
First mover advantage, 358
Fisheye views, 255–257
Flash cookie, 55
Flickr—Sharing your photos, 380
Focused crawlers, 83
Focused mobile search, 299–301
Focused subgraph, 118
FolkRank, 387
Folksonomies, 402
Folksonomy, 383–384
Follower spam, 352
Followers, 351
Forward buttons, 214–215
Free riding problem, 329
Friend search, 372
Friendster (www.friendster.com), 317
F-score, 137
Full-text retrieval systems, 29
Fusion algorithms, 169–170
Future of web search and navigation, 419–423
G
Generalized first price (GFP) auction, 162
Generalized second price (GSP), 163
Geodesic path, 321
Geographic information retrieval, 303
Geoparsing, 303
Get, 328
Getting lost in hyperspace, 39–42
Giant component, 366
Global bridge, 321
Global efficiency, 365
Global search service, requirements, 85–88
   BigTable database system, 87
   fault-tolerant software, 85
   Google File System (GFS), 87
   Google query, 86
   MapReduce algorithm, 87
   query execution, phases, 86
   Gnutella network, 328–329
   Golden triangle, 139
   Google (www.google.com), 8, 68–72
   Google AdSense (www.google.com/adsense) program, 162
   Google Analytics (www.google.com/analytics), 235
   Googlearchy, 123
   Google bomb, 124, 144
   Google dance period, 81
   Google death penalty, 66
   Google File System (GFS), 87
   Google Hacks, 61
   Google Latitude (www.google.com/latitude), 278
   Google, reach of, 16
      in China, 16
      in Far East, 16
      Germany, 16
Google, reach of (Continued)
in Korea, 16
Japan, 16
in United States, 16
Google squared
(www.google.com/squared), 191
Google Trends (www.google.com/trends), 78
Google-Watch (www.google-watch.org), 67
Google-Watch
(www.google-watch-watch.org), 111
H
Hadoop (http://hadoop.apache.org), 87
HealthMap (www.healthmap.org), 396
Hidden text, 97
Hidden web, 10
Highly optimized tolerance (HOT), web evolution via, 360–361
History enriched navigation, 369
History list, 219
Hit and miss (HM), 242
Home button, 213
Home page detection, 107
Hot list, See Bookmarks tool
HousingMaps (www.housingmaps.com), 396
Hubs, 332
Hybrid P2P networks, 330–331
superpeers, 330
Hypercard programming environment, 224–225
buttons, 224
card, 224
stack, 224
Hyperlink-Induced Topic Search (HITS), 93, 117–120
base set, 117
focused subgraph, 118
root set, 117
topic drift, 119
Hyperlinks, 211–212
Hypertext markup language (HTML), 4, 211
structure weighting, 104–105
Hypertext orientation tools, 223
Hypertext transfer protocol (HTTP), 4
Hypertext, history of, 3–6
I
Image search, 194–200
CAPTCHA, 199–200
content-based image search, 196–198
content-based retrieval, 194
for finding location-based information, 200–201
pseudorelevance feedback, 197
reCAPTCHA, 199–200
relevance feedback, 197
text-based image search, 195–196
VisualRank, 198–199
I-mode Service, 275–277
location-aware services, 276
short message service (SMS), 276
Implicit feedback, 278, 335
Incentives in P2P systems, 332–333
Incoming links, counting, 122–123
Indexer, 79–80
inverted file, 79
posting list, 79
source URL, 79
Indirect greedy algorithm, 377
Information extraction, 244–245
textual, 244
Information gatekeepers of web, 64–66
search engines as, 64–66
Information presentation on mobile device, 287–291
Information retrieval (IR) systems, 6
web search and traditional, comparing, 29–32
precision, 30–32
recall, 30–32
Information scent absorption rate, 229
Information scent notion, 229
Information seeking
on mobile devices, 284
on web, 24–28
direct navigation, 24–25
problems with, 27–28
Informational links, 109–110
Informational query, 28
InfoSeek, 7
Initial Public Offering (IPO), 69
Inktomi, 7
Inquirus, 172
Instant messaging social network, 314
Interface, search, 81
Internet Protocol (IP) address, 14
Internet service provider (ISP), 168
Internet, power-law distributions in, 355
Inverse document frequency (IDF), 92, 98–100
Inverse fisheye heuristic, 242
Inverted file, 79
Invisible web, 10
iPhone (www.apple.com/iphone), 283
Item-based collaborative filtering, 337

J
Junk e-mail, 80
JXTA P2P search, 332

K
k means clustering algorithm, 217
Kartoo, 172, 258, 261
Keystrokes per character (KSPC), 286
Keyword TF–IDF values, computing, 100–102
Keywords, search using, 77–78
KidsClick (www.kidsclick.org), 203
k-nearest neighbors, 336
Known item searches, 107
Known item searches, 32

L
Labsheet10, 99
Laid back mobile search, 300
Landmark page, 41–42
Law of Participation, 355–357
Law of Surfing, 355–357
Learning resources, delivery of, 281–282
Learning to rank, 133–134
Lesstap, 285
Limewire (www.limewire.com), 329
Linearity theorem, 184
Link-based metrics, 108–130
citation analysis, 127–129
Google-Watch
(www.google-watch-watch.org), 111
heuristics, 126
HITS, 117–120, See also Hyperlink-Induced Topic Search (HITS)
importance, 110–112
incoming links, counting, 122–123
informational links, 109–110
link analysis and content relevance, 110
link spam, 125–127
PageRank bias against New Pages, 123
PR Ad Network, 111
referential links, 109–110
Link marker, 213
Link text (or anchor text), 39, 80, 103–104, 213
LinkedIn (www.linkedin.com), 318
Local bridge, 321
Local efficiency, 365
Local site search versus global web search, 32–34
Locally envy-free equilibrium, 163
Location-aware mobile search, 303–305
Location-aware searches, 276, 298
predefined search for objects that the user may pass, 299
real-time searching for nearby objects, 298
Location awareness, 295
Location-based information, image search for, 200
Location-based services (LBSs), 277
Location breadcrumbs, 221
Logarithmic binning, 354
Log–log plot, 354
Loopt (www.loompt.com), 278
Lotka’s law, 353
Love Bug virus, 367

M
Machine assistance, 42–46
in user search and navigation, 42–46
algorithms, 42–43
naive Bayes classifier, 43–46
Machine learning algorithms testing, blogs for, 349
Manhattan distance, 375
Mapping cyberspace, 262
MapQuest (http://wireless.mapquest.com), 278
MapReduce algorithm, 87
Maps, 223
Markov chains, 49–54, 238–242
branching processes, 50
dynamic Markov modeling, 241
and Markov property, 49–50
and the probabilities of following links, 50–52
queueing theory, 50
random walks, 50
Markov chains (Continued)
and relevance of links, 52–54
simulation using, 50
state, 50
suffix tree, 241
Mashups, 396–397
Maximum likelihood estimate, 158, 242
M-Commerce, 277–278
Mean absolute error (MAE), 242, 340
Mean reciprocal rank (MRR), 137
MediaWiki (www.mediawiki.org), 399
Memes, 349–350
Memex concept, 3–4
adaptive memex, 4
MetaCrawler (www.metacrawler.com), 171
Metasearch, 168–178
Borda count, 170
classifying search results, 175–178
snippet enriched query, 177
support vector machine (SVM) classifier, 177
clustering search results, 173–175
Condorcet count, 170
fusion algorithms, 169–170
operational metasearch engines, 170–173
Dogpile (www.dogpile.com), 170
Inquirus, 172
Kartoo, 172
MetaCrawler (www.metacrawler.com), 171
ProFusion, 171
SavvySearch (www.savvysearch.com), 171
Vivisimo (www.vivisimo.com), 172
rank aggregation algorithms, 169
rank fusion, 168
weighted HITS, 170
Microblogging, 350–352
Microconversions, 238
Milgram’s small-world experiment, 312–313
Mobile computing, paradigm of, 273–277
i-mode service, 275–277
wireless markup language (WML), 274
Mobile device interfaces, 282–291
Amazon’s Kindle (www.amazon.com/kindle), 283
Android (www.android.com), 283
BlackBerry (www.blackberry.com), 283
Fennec (https://wiki.mozilla.org/Fennec), 283
information presentation on, 287–291
iPhone (www.apple.com/iphone), 283
mobile web browsers, 282–283
NetFront (www.access-company.com), 283
Opera Mobile (www.opera.com/mobile), 283
Palm (www.palm.com), 283
Safari (www.apple.com/iphone-3gs/safari.html), 283
Skyfire (www.skyfire.com), 283
Symbian (www.symbian.org), 283
WebKit (www.webkit.org), 283
Windows Mobile (www.microsoft.com/windowsmobile), 283
Mobile query log analysis, 301–302
Mobile search, 295–306
candidate set, 300
focused mobile search, 299–300
gеographic information retrieval, 303
interfaces, 296–298
laid back mobile search, 300
location-aware mobile search, 295
location-aware searches, 298
mobile query log analysis, 301–302
mobile search interfaces, 296–298
personalization, 295, 302–303
search engine support for, 298–299
WithAir system, 298
Mobile web/Mobile web services, 277–282
information seeking on, 284
navigation problem in, 291–295
adaptive mobile portals, 292–294
adaptive web navigation, 294–295
click-distance, 291–292
portals, 277
text entry on, 284–286
voice recognition for, 286–287
Model-based collaborative filtering, 338–339
Monotone, 377
Monte Carlo methods in PageRank computation, 116–117
MovieLens (http://movielens.umn.edu), 346
Mozilla Geode
(https://wiki.mozilla.org/Labs/Geode), 278
MrTaggy (http://mrtaggy.com), 387
MSN Search, 71
Multiplicative process, web evolution as, 359–360
Multitap, 285
Museum experience recorder, 264–265
Myspace (www.myspace.com), 318

Naive Bayes (NB) classifier, 280, 338
for web pages categorizing, 43–46
assumption, 46
Bayesian networks, 44
belief networks, 44
Narrowband connection, 17
Natural language annotations, 187–190
Naver (www.naver.com), 16
Navigation, 38–58, 209–269, See also
Breadcrumb navigation; Browsing;
Problem of web navigation
frustration in, 211–213
404 error—web page not found, 212
HTML, 211
hyperlinks, 211–212
pop-up ads, 212
referential integrity problem, 212
surfing, 211–212
web site design and usability, 211–213
machine assistance in, 42–46
in physical spaces, 262–266
problem in mobile portals, 291–295,
See also under Mobile web
navigation tools, 213–225, See also
Bookmarks tool
back buttons, 214–215
basic browser tools, 213–214
click here, 213
home button, 213
link marker, 213
link text (or anchor text), 213
tabbed browsing, 213
built-in to the browser, 214
forward buttons, 214–215
hypercard programming environment,
224–225
quicklinks, 222–223
revisitation rate, 215
Navigation using search engine, 26–27
navigation (or surfing), 27
query formulation, 26
query modification, 27
selection, 26
Navigational metrics, 225–230
centrality notion, 226
connectivity notion, 226
information scent absorption rate, 229
information scent notion, 229
potential gain, 226–227
usability, measuring, 229–230
web user flow by information scent, 229
Navigational query, 28, 159
Navigation-oriented sessioning, 237
Nav-search, 2491250
Nearest-neighbor (NN) classification
algorithm, 280
Neighbor-of-neighbor greedy (NoN-greedy)
algorithm, 376
Netcraft (www.netcraft.com), 15
Netflix (www.netflix.com), 342–346
Cinematch, 342
probe subset, 343
quiz subset, 343
root mean squared error (RMSE), 342
test subset, 343
NetFront (www.access-company.com), 283
Netscape, 56
New Pages, PageRank bias against, 123
Newsmap (www.newsmap.jp), 397
Nielsen’s hypertext, 224–225
Nonbinary relevance assessment, 94
Non-English queries, 106
Normal distribution, 352
Normalized discounted cumulative gain
(NDCG), 138
O
Online computation of pagerank, 116
Online system (NLS), 4–5
Online trading community, 407–412, See
also eBay
Open APIs, 396–398
Open Directory, 151
Open domain question answering, 192–193
Opera Mobile (www.opera.com/mobile), 283
Opinion mining, 390–392
  comparative sentence and relation extraction, 393
  feature-based opinion mining, 391–392
  sentiment classification, 392–393
Outride’s personalized search, 186–187
Overture (www.overture.com), 8
Overview diagrams, 223, 253–255

P
Page Hunt, 139
Page swapping, 97
PageRank, 110–116, 225
  bias against new pages, 123
  Monte Carlo methods in, 116–117
  online computation of, 116
  personalized PageRank, 184–186
    BlockRank, 185
    query-dependent PageRank, 185
    topic sensitive, 185
  popularity of, 129–130
  sculpting, 126
  Weblogs influence on, 124–125
  within a community, 123–124
Paid inclusion, 152–153
Paid placement, 68, 153–157
Pajek, large network analysis software, 326
Palm (www.palm.com), 283
Parakeet, 286
Path breadcrumbs, 221
Pattie Maes, 347
Pay per action (PPA) model, 149, 165–166
Pay per click (PPC), 149, 154
Pay per impression (PPM), 153
Peer-to-peer (P2P) networks, 326–333, See also Hybrid P2P networks
  BitTorrent protocol, 329
  centralized, 327–328
  decentralized, 328–330
    free riding problem, 329
    get, 328
  Gnutella, 328
  ping, 328
  pong, 328
  push, 328
  query response, 328
Gnutella network, 329
  incentives in, 332–333
  limewire (www.limewire.com), 329
Permalinks, 348
Personalization, 178–187, 295
  approaches to, 179
    click-based approach, 179
    topic-based approach, 179
  blind feedback, 183
  of mobile search, 295, 302–303
  outride’s personalized search, 186–187
  personalized PageRank, 184–186
    BlockRank, 185
    query-dependent PageRank, 185
    topic sensitive, 185
Personalized Results Tool (PResTo!), 180–181
  privacy, 182
  pseudorelevance feedback, 183
  relevance feedback, 182–184
  scalability, 182
  versus customization, 180
  web usage mining for, 246
Personalized news, delivery of, 278–281
  Daily Learner, 279
  explicit feedback, 278
  implicit feedback, 278
  long-term user model, 280
  short-term user model, 280
  WebClipping2, 281
Phrase matching, 102
Physical space, navigation in, 262–266
PicASHOW, 196
Ping, 328
PocketLens, 346
Pong, 328
Popularity, 54
Popularity-based metrics, 130–135
  BrowseRank, 134–135
  continuous Markov chain, 135
  Direct Hit, 130–132
  document space modification, 132
  learning to rank, 133–134
  query log data to improve search, 132–133
  RankNet, 134
  TrustRank, 135
Pop-up ads, 212
Portal menu after personalization, 294
Portals, mobile, 277
Porter stemmer, 95
Potential gain of web page, 226–227
Power Browser project, 287–289
Power buyers, 412
Power-law distributions in web, 352–368
detecting, 353–355
in internet, 355
Law of Participation, 355–357
Law of Surfing, 355–357
Power laws, 99, 154
Power method, 114–115
Power sellers, 412
PR Ad Network, 111
Precision, 30–32, 137
recall versus, 31
Prediction market, 400
Prediction method, 243
Preferential attachment, web evolution via,
357–359
Principal component analysis (PCA), 265
Privacy, 182
Probabilistic best first search algorithm, 248
Probe subset, 343
Problem of web navigation, 38–58
cookies, 55
getting lost in hyperspace, 39–42
semantics of web site and business
model, conflict between, 57–58
web site owner and visitor, conflict
between, 54–57
Problems of search, web and, 9–37
tabular data versus web data, 18–20
usage statistics, web, See Usage
statistics, web
web site size statistics, See Size statistics,
web
ProFusion, 171
Proximity matching, 76
Proxy bidding, 408
Pseudoratings, 339
Pseudorelevance feedback, 183, 197–198
Publisher click inflation, 166
Push, 328
Queries, 28–29
formulation, navigation, 26
informational, 28
interpreting, 96
modification, 27
navigational, 28
query-dependent PageRank, 185
response, 328
suggestions, 107–108
transactional, 28
Query engine, 80–81
junk e-mail, 80
search interface, 81
Query logs, search engine, 73–78
data to improve search, 132–133
English-based queries, 74
keywords search, 77–78
proximity matching, 76
query index, 78
query syntax, 75–76
temporal analysis, 74
topical analysis, 74
Question answering (Q&A) on web,
187–194
Ask Jeeves’ Q&A technology, 188
factual queries, 190–192
Google squared
(www.google.com/squared), 191
Mulder interface, 193
named entity recognition, 192–193
natural language annotations, 187–190
open domain question answering,
192–193
redundancy-based method, 192
semantic headers, 193–194
Wolfram Alpha
(www.wolframalpha.com), 190–192
Quicklinks, 222–223
Quiz subset, 343
QWERTY keyboard, 286
R
Radio frequency identification (RFID), 422
Rank aggregation algorithms, 169
Rank by bid, 163
Rank by revenue, 163
Rank fusion, 168
Rank sink, 112–113
Ranking algorithms, inferring, 142–143
RankNet, 134
Rapid serial visual presentation (RSVP),
257–258
Readers, 223
Reality analytics, 266
context dimension, 266
Reality analytics (Continued)
   social dimension, 266
   spatial dimension, 266
   temporal dimension, 266
Reality mining, 265
Really simple syndication (RSS), 244
Real-timeweb, 350–352
Real world, navigating in, 265–266
Real-world web usage mining, 262–264
Recall, 30–32, 137
   versus precision, 31
reCAPTCHA, 200–201
Redirects, 97
Redundancy-based method, 192
Referential integrity problem, 212
Referential links, 109–110
Refreshing web pages, 84
Regularization, 344
Related concepts, 107–108
Relation, 320
Relation extraction, 393
Relationship structures, 244
Relevance, 94–108, 136, See also Content relevance
   Relevance feedback, 182–184, 197
Repeat order rate, 234
Repeat visitor rate, 234
Revisitation rate, 215
Rewiring probability, 363
Ridge regression, 345
Ringo, 347
Robots exclusion protocol, 84–85
Root mean squared error (RMSE), 342
Rule-based systems, 245
S
SaaS, 398
Safari (www.apple.com/iphone/iphone-3gs/safari.html), 283
SavvySearch (www.savvysearch.com), 171
Scalability of CF, 182, 341
Scale-free network
   robustness of, 366–368
   threshold value, 367
   vulnerability of, 366–368
Scale-invariant feature transform (SIFT) algorithm, 198
Scent of information, 229
Search and navigation, difference between, 34–35
Search engine optimization (SEO), 91, 98
Search engine support for mobile devices, 298–299
Search engine wars, 68–72
   Ask Jeeves (www.ask.com), 72
   Bing (www.bing.com), 70–72
   Cuil (www.cuil.com), 72
   Google (www.google.com), 68–72
   MSN search, 71
   web crawler, 71
   Yahoo (www.yahoo.com), 70
Search engines, See also individual entries
   history of, 6–8
Searching the web, 60–89
   mechanics of, 61–64
   optimization, 66
   visibility, 66
Seed, 331
Selection, in navigation, 26
Self-information, 100
Semantic gap, 196
Semantic headers, 193–194
Semantics of web site and business model, conflict between, 57–58
Sentiment classification, 392–393
Sessionizing, 237
   navigation-oriented, 237
   time-oriented, 237
Sharing scholarly references, 383
Shill account, 410
Short message service (SMS), 276
Shrink-wrapped software (SWS), 398
Similarity matrix, 173
Singular value decomposition (SVD), 174, 345
Sitemaps (www.sitemaps.org), 153
‘Six degrees of separation’ notion, 23
Size statistics, web, 10–15
   capture-recapture method to measure, 12
   coverage issue, 11
   Cuil (www.cuil.com), 13
   deep web data, 10–11
   hidden or invisible web, 10
   IP (Internet Protocol) address, 14
   Netcraft (www.netcraft.com), 15
   Skyfire (www.skyfire.com), 283
   Skype, 407
   Skyscraper ads, 154
Small-world networks, 361–366
  navigation within, 375–379
  awareness set, 377
  indirect greedy algorithm, 377
  monotone, 377
  random, 364
  regular, 364
  small world, 364
Small-world structure of web, 23–24
Sniping, 409–410
Snippet enriched query, 177
Social aggregator, 351
Social data analysis, 259–262
Social dimension of reality analytics, 266
Social network analysis, 320–326
Social network start-ups, 316–320
  Dodgeball, 319
  Facebook (www.facebook.com), 318
  Friendster (www.friendster.com), 317
  Linkedin (www.linkedin.com), 318
  Myspace (www.myspace.com), 318
  Visible Path, 319
Social networks, 309–417
  centrality, 322–324
  collaboration graphs, 313–314
  description, 311–320
  Erdös collaboration graph, 313
  instant messaging social network, 314
  Milgram’s small-world experiment, 312–313
  navigation strategies testing in, 379
  e-mail network, 379
  online student network, 379
  searching in, 369–379
  history enriched navigation, 369
  P2P networks, 374
  social search engines, 370–373
  social web, 314–316
terminology, 320–322
  clustering coefficient, 321
  degree distribution, 321
  geodesic, 321
  global bridge, 321
  local bridge, 321
  relation, 320
  strength of weak ties, 322
  Ties connect, 320
Social search engines, 370–373
  collaborative search, 372
  collective search, 372
  community search, 372
  friend search, 372
Social tagging, 379–389
  classifying tags, 389
  clustering tags, 389
  efficiency of tagging, 388
  Flickr—Sharing your photos, 380
  YouTube (www.youtube.com), 380
Software as a service, 398
Source URL, 79
Spam, link spam, 125–127
Spatial dimension of reality analytics, 266
Special purpose search engines, 200–205
  Amazon.com (www.amazon.com), 201
  arXiv.org e-Print archive (http://arxiv.org), 202
  CiteSeer/CiteSeerX, 202
  Digital Bibliography and Library Project (DBLP), 202
  FindLaw (www.findlaw.com), 202
  KidsClick (www.kidsclick.org), 203
Speech recognition for mobile devices, 286–287
Spell checking, 105–106
Spider traps, 85
Sponsored search auctions, 161–165
  generalized first price (GFP) auction, 162
  generalized second price (GSP), 163
  Google AdSense (www.google.com/adsense) program, 162
  locally envy-free equilibrium, 163
  rank by bid, 163
  rank by revenue, 163
Sponsored search, 153–157
Stack, 224
Stack-based back button, 214
StarTree, 257
State network, 50
Stemming technique, 95
  Porter stemmer, 95
Stochastic approach for link-structure analysis (SALSA), 93, 120–122
Stop words, 77
Stratum, 228
Strongly connected component (SCC), 22, 228
Structural analysis of web site, 228
  compactness, 228
  stratum, 228
Structural equivalence, 324
Structural hole, 322
Structure mining, 231
Structure of web, 20–24
  bow-tie structure, 21–23
  'six degrees of separation' notion, 23
  small-world structure, 23–24
  strongly connected component (SCC), 22
Structured Query Language (SQL), 19
Stuffing, 97
StumbleUpon (www.stumbleupon.com), 372–373, 387
Suffix tree clustering (STC), 174
Suffix tree, 241
Superpeers, 330
Supplementary analyses, 237–238
Support vector machine (SVM) classifier, 177
Surfer, identifying, 236–237
  by the host name or IP address, 236
  through cookie, 237
  through user login to the site, 237
Surfing, See Navigation
Symbian (www.symbian.org), 283
Syndication, 395–396
Synonyms, 102–103
System for the mechanical analysis and retrieval of text (SMART), 6

T
T9 predictive text method, 285–286
Tabbed browsing, 213
Tabular data versus web data, 18–20
  SQL (Structured Query Language), 19
Tag clouds, 384–385
Tag search, 385–388
Tagging web pages, 303
Take rate, 234
Technorati (www.technorati.com), 348, 387
Teleportation, 112
Temporal difference learning, 35
Temporal dimension of reality analytics, 266
Term frequency (TF), 91, 96–99
  cloaking, 97
  computation of, 97
  doorway pages, 97
  duplicate pages, 97
  hidden text, 97
  page swapping, 97
  redirects, 97
  stuffing, 97
term frequency-inverse document frequency (TF-IDF), 92, 96
tiny text, 97
Test set, 243
Test subset, 343
Text-based image search, 195–196
  image containers, 196
  image hubs, 196
  PicASHOW, 196
Text entry on mobile devices, 284–286
  lesstap, 285
  multitap, 285
  T9 predictive text method, 285–286
two-thumb chording, 285
Text REtrieval Conference (TREC) competitions, 141
Textual information extraction, 244
Threshold value, 367
Ties connect, 320
Tightly knit community (TKC), 121
Time-oriented sessionizing, 237
Time-to-live (TTL) number, 328
Timway (www.timway.com), 74
Tiny text, 97
Toolbars, search engine, 215–216
Topic drift, 119
Topic locality, 110
Topic-based approach, 179
Top-m terms, 108
Top-n precision, 108, 137
Trails, 46–49, 249–250, See also Best trail algorithm
  creating, 47
  authored, 47
  derived, 47
  emergent trails, 47
  as first class objects, 46–49
Transactional query, 28
Trawling, 325
Trend graphs, 349
Triad formation, 365
True positives, 402
Truncation, 96
TrustRank, 126, 135
Twitter (www.twitter.com), 351
  as a powerful 'word of mouth' marketing tool, 352
Two-thumb chording, 285
INDEX 477

U
Unified resource locator (URL) to identify web pages, 4
Universal predictors, 243
URL analysis, 104
Usability of web sites, measuring, 229–230
Usage mining, 231
Usage statistics, web, 15–18
User-based collaborative filtering, 335–337
  explicit rating, 335
  implicit feedback, 335
  vector similarity, 335
User behavior, 158–160
User search, machine assistance in, 42–46

V
Vector, 108
Vector similarity, 198, 335
VeriTest (www.veritest.com), 138
Viewing graph, 245
Views, 420
Virtual space, navigation in, 262–266
Visible Path, 319
Visitors, 54–57, See also Web site owner and visitor, conflict between
Visual links, 198
Visual search engines, 251, 258–259
Visualization that aids navigation, 252–262
  hierarchical site map, 255
  navigation patterns, visualizing, 252–253
  online journal application, 256
  Open Directory home page, 254
  overview diagrams, 253–255
  query-specific site map, 256
  web site maps, 253–255
  WebBrain (www.webbrain.com), 253
Visualizing trails within a web site, 257–258
VisualRank, 198–199
Vivisimo (www.vivisimo.com), 172
Voice recognition for mobile devices, 286–287
Vox Populi (voice of the people), 132

W
Weatherbonk (www.weatherbonk.com), 396
Web
  conventions, 4
  HTML, 4
HTTP, 4
URL, 4
  history of, 3–6
Web2.0, 393–399
  Ajax, 394
Craiglist (www.craigslist.org), 396
HealthMap (www.healthmap.org), 396
HousingMaps (www.housingmaps.com), 396
Mashups, 396–397
Newsmap (www.newsmap.jp), 396
open APIs, 396–398
software as a service, 398
syndication, 395–396
Weatherbonk (www.weatherbonk.com), 396
Widgets, 396–397
Web analytics, 233–235
Web communities, 324–326
  bipartite graph notion, 325
  block, 324
  blockmodeling, 324
  clustering, 324
  Pajek, large network analysis software, 326
  structural equivalence, 324
  trawling, 325
Web crawler, 71, 79, See also Crawling the web
Web evolution
  as a multiplicative process, 359–360
  via preferential attachment, 357–359
Web history, 187
Web navigation, See Navigation
Web ontology language (OWL), 420
Web pages, identifying, 219–221
  by their title, 219
  by their URL, 219
  by a thumbnail image of page, 219–220
Web pages, processing, 94–96
Web portal, 25
Web search, 32
Web site maps, 253–255
Web site owner and visitor, conflict between, 54–57
  objectives, 55
Web usage mining, 52
  applications of, 242–244
  for personalization, 246
WebBrain (www.webbrain.com), 253
WebFountain, 243
WebKit (www.webkit.org), 283
Weblog file analyzers, 235–236
Weblogs influence on PageRank, 124–125
WebTwig, 297
Weighted HITS, 170
Weka (www.cs.waikato.ac.nz/ml/weka), 402
Widgets, 396–397
Wiki, 399
   MediaWiki (www.mediawiki.org), 399
Wikipedia, 402–406
Windows Mobile (www.microsoft.com/windowsmobile), 283
Wireless markup language (WML), 274
WithAir system, 298
Wolfram Alpha (www.wolframalpha.com), 190–193
Wordle (www.wordle.net), 384, 386
Working of search engine, 91–145,
   See also Content relevance; Link-based metrics; Popularity-based metrics
Wrapper induction, 245
Wrapper, 274
X
Xanadu, 4–5
Y
Yahoo (www.yahoo.com), 7, 70
Yandex (www.yandex.com), 16
YouTube (www.youtube.com), 380
Z
Zipf’s law, 99, 353