INDEX

*a priori* algorithm, 201–204
Association rules for web usage mining, 197–204
Authorities and hubs, 9, 53–55
Average time per page, 183–185

Background knowledge, 133
Basket transformation, 171–183
Bayes rule, 76, 125
Between-cluster variation, 194
Binning, 199–201
BIRCH clustering algorithm, 193–197
Boolean representation, 16, 21

C4.5 algorithm, 208–210
Classes-to-clusters evaluation, 106–108
Classification, 115–139
  feature selection, 121–125
    entropy, 121–122
    InfoGain, 123–124
    information gain, 122
    similarity-based feature selection, 122
  general setting and evaluation techniques, 115–117
  cross-validation (CV), 117
  holdout approach, 117
  supervised learning framework, 115–116
  naive Bayes algorithm, 125–131
  Bayes rule, 125
  Laplace estimator, 128
  naive Bayes assumption, 125
  nearest-neighbor algorithm, 118–120
    *k*-nearest-neighbor (*k*-NN), 119
    distance-weighted, 120
    one-nearest-neighbor (1-NN), 118
  numerical approaches, 131–133
  linear combination, 131

Classification and regression trees (CART), 204–208
  branches, 204
  decision nodes, 204
  decision rules, 206, 208
  decision tree, 204
  GINI index, 205
  leaf node, 204
  root node, 204
  root node split, 206
Classification for web usage mining, 204–210. *See also* Web usage mining
Classification problem, 76–78
Clementine, xiv
Clickstream analysis, 147–148
Closed world assumption, 134
Clustering, 61–88
  agglomerative algorithm, 64
  collaborative filtering (recommendor systems), 84–85
  definition of, 193–194
  evaluating, *see* Evaluating clustering
  farthest-neighbor, 64
  hierarchical agglomerative, 63–69
  agglomerative algorithm, 64
  cosine similarity, 63
dendrogram, 63
farthest-neighbor, 64
nearest-neighbor, 64
Clustering (Continued)
similarity
   average, 64
   between cluster centroids, 64
   intracluster, 64, 68
   minimum, 64
k-means, 69–73
   k-means algorithm, 69
   minimum variance, 70
   minimum variance, 70
   nearest-neighbor, 64
probability-based, 73–84
   Bayes rule, 76
   classification problem, 76–78
   clustering problem, 78–84
   expectation maximization (EM) algorithm, 79
finite mixture problem, 74–75
inddependence assumption, 77
labeled data set, 74
log-likelihood criterion function, 80
mean, class, 75
naive Bayes, 77
optimization, 79
probability density function, 76
probability of sampling, 75
standard deviation, class, 75
problem, 78–84
for web usage mining, 193–197
Collaborative filtering, 84–85
Common log format, 151
Companion website, xiv
Confidence difference method, 202
Cookies, 164
Cosine similarity, 24, 25, 36–38, 63
Crawling the web, 6–12
   address resolution, 11
   authority, 9
   breadth first crawling, 8, 10
   depth first crawling, 8, 9
   guard, 11
   guided search, 12
   html, 6
   hub, 8
   robot exclusion protocol, 12
   text repository, 11
   uninformed graph search, 12
   url format, 6
web archive, 12
web basics, 6
web crawlers, 7–12
web as directed graph, 7
webSPHINX, 8–10
Cross-industry standard process for data mining (CRISP-DM), 144–147
business (research) understanding phase, 145
data understanding phase, 145
data preparation phase, 145
deployment phase, 146
evaluation phase, 145
input stage, 146
modeling phase, 145
pattern analysis stage, 146
pattern discovery stage, 146
preprocessing stage, 146
Cross-validation (CV), 117
Data cleaning and filtering, 158–162
Data compression, 104
Data mining, xi
Data Mining the Web: Uncovering Patterns in Web Content, Structure and Usage
how book is structured, xi
why book is needed, xii
white-box approach, xii
as textbook, xv
Decision rules, 206, 208
Decision tree, 204
Dendrogram, 63
De-spidering the web log file, 163–164
Directories, 171–173
Document ranking, 23–26
Document representation, 15–18
Document resemblance, 41–42
Entropy, 111–112, 121–122, 208
Entropy reduction, 208
Euclidian distance, 24
Evaluating clustering, 89–114
   classes-to-clusters evaluation, 106–108
   error-based attribute evaluation, 107
   OneR, 107
   entropy, 111–112
MDL-based model and feature evaluation, 100–106. See also Minimum description length (MDL) model
precision, recall, and $F$-measure, 108–111
   confusion matrix, 108
contingency table, 108
error cost, 108
false negative, 108
false positive, 108
F-measure, 110
harmonic mean, 111
precision, 109
recall, 109
true negative, 108
true positive, 108
probabilistic criterion functions, 95–100.
   See also Probabilistic criterion functions
similarity-based criterion functions, 90–95. See also Similarity-based criterion functions
Evaluating search quality, 32–35
   precision, 32
   average, 33
   interpolated, 34
recall, 32
Expectation maximization (EM) algorithm, 79
Exploratory data analysis (EDA), 177
   average time per page, 183–185
   duration for individual pages, 185–187
   page duration calculation procedure, 186
   number of visit actions, 177–178
   page requests, 177
   visit actions, 177
   relationship between visit actions and session duration, 181–183
   regression analysis, 181
   regression equation, 182
   slope, 182
   session duration, 178–181
   session duration calculation procedure, 179
   visit time, 179
for web usage mining, 177–190
Extended common log format, 151–153
F-measure, 110
Feature selection, 105–106, 121–125
Finite mixture problem, 74–75
Flag variable, 172
Generalized rule induction (GRI) algorithm, 199
GINI index, 205
Holdout approach, 117
Hyperlink-based ranking, 47–58
   authorities and hubs, 53–55
      topic distillation, 53
   enhanced techniques for page ranking, 56–57
      nepotism links, 56
      outliers, 57
      topic drift, 56
      topic generalization, 56
link-based similarity search, 55–56
   pagerank, 50–53
      pagerank algorithm, 52
      pagerank score, 50
      random walk, 50
social networks analysis, 48–49
   eigenvalue, 48
   eigenvector, 48
   fixed point, 48
   power iteration method, 49
   prestige, 48
Independence assumption, 77
Indexing and keyword search, 13–32
   advanced text search, 28–29
   anchor tag, 31
   anchor text, 30, 31
   approximate string matching, 29
   bag-of-words representation, 17
   Boolean representation, 16, 21
   B-trees, 19
   directory page, sample, 14
   document ranking, 23–26
   document representation, 15–18
   dot product, 24
   Euclidian distance, 24
   feature selection, 26
   formal representation, 16
   hash tables, 19
   headings, 30
   implementation considerations, 19
   information retrieval (IR), 13
   inverse document frequency (IDF), 22
   inverted index, 17
   keyword search, 13
   metatags, 30
   metric function, 24
   n-grams, 29, 42
   parameters, 16
   part-of-speech tagging, 17
Indexing and keyword search (Continued)
phrase dictionary, 29
phrase search, 28–29
proximity measure, 23–24
pseudorelevance feedback, 27
query time, 20
query vector, 23–24
relevance feedback, 26–28
relevance ranking, 13, 20
Rocchio’s method, 27
structured data, 13
Structured Query Language (SQL), 13
tagging, 29
terms, 16
term frequency (TF), 17, 21
term-document matrix, 17
term-document matrix examples, 18, 19
text corpus, 16
TFIDF, 22
tokenizing documents, 15
vector space model, 21–23
using html structure in keyword search, 30–31
web document, sample, 15
indicator variable, 172
InfoGain, 123–124
Information gain, 122, 208
Information retrieval and web search, 1–46
Crawling the web, 6–12. See also
Crawling the web
evaluating search quality, 32–35. See also
Evaluating search quality
indexing and keyword search, 13–32. See also
Indexing and keyword search
similarity search, 36–42. See also
Similarity search
web challenges, 3–5
semantic web, 5
topic directories, 5
web growth, 3
web search engines, 4
Jaccard similarity, 38–41
$k$-nearest-neighbor ($k$-NN), 119
distance-weighted, 120
Laplace estimator, 128
Link-based similarity search, 55–56
Log-likelihood criterion function, 80
Linearly separable, 132
Microsoft IIS log format, 153–154
Min-max normalization, 193
Minimum descriptive length (MDL) model, 100–106
data compression, 104
feature selection, 105–106
generalization by dropping conditions, 100
MDL-based model evaluation, 102–105
minimum description length principle, 101–102
Occam’s razor, 101
Modeling for web usage mining, 191–212
affinity analysis and the $a$ priori
algorithm, 197–199
affinity analysis, 197
antecedent, 197
association rule, 198
confidence, 198, 199
consequent, 197
generalized rule induction (GRI) algorithm, 199
market basket analysis, 197
support, 198, 199
applying the $a$ priori algorithm to the
CCSU web log data, 201–204
confidence difference method, 202
model, 203
pattern, 203
posterior probability, 202
prior probability, 202
BIRCH clustering algorithm, 193–197
C4.5 algorithm, 208–210
confluence of evidence, 210
convergence of models, 210
entropy, 208
entropy reduction, 208
information gain, 208
classification and regression trees
(CART), 204–208. See also
Classification and regression trees
clustering, definition of, 193–194
between-cluster variation, 194
min-max normalization, 193
within-cluster variation, 194
z-score standardization, 194
discretizing the numerical variables, 199–201
binning, 199
methodology, 192–193
test set, 193
training set, 192
validation set, 193
Modeling methodology, 192–193

n-grams, 29, 42
Naive Bayes algorithm, 77, 125–131
Naive Bayes assumption, 125
Nearest-neighbor algorithm, 118–120
Numerical approaches, 131–133

Occam’s razor, 101
One-nearest-neighbor (1-NN), 118

Page duration, 185–187
Page extension, exploration, and filtering, 161–162
Page requests, 177
Pagerank, 50–53
Path completion, 170
Posterior probability, 202
Power iteration method, 49
Precision, 32, 109
Preprocessing for web usage mining, 156–176
data cleaning and filtering, 158–162
page extension, exploration and filtering, 161–162
time stamp, creating, 159
variable extraction, 159
de-spidering the web log file, 163–164
crawlerbot, 163
directories and the basket transformation, 171–173
basket transformation, 172
flag variable, 172
indicator variable, 172
further data preprocessing steps, 174
need for, 156–158
path completion, 170
session identification, 167–170
reference length approach, 167
session identification procedure, 169
site visit, 167
time delay, 167
user session, 167
user identification, 164–167
cookies, 164
user identification procedure, 167
Prestige, 48

Prior probability, 202
Probability density function, 76
Probabilistic criterion functions, 95–100
category utility, 96
probability matching strategy, 97
Cobweb clustering algorithm, 97
Proximity measure, 23–24

Query vector, 23–24

Recall, 32, 109
Recommender systems, 84–85
Reference length approach, 167
Regression analysis, 181
Relational learning, 133–137
Relevance feedback, 26–28
Relevance ranking, 13, 20
Rocchio’s method, 27

Session duration, 178–181
Session identification, 167–170
Similarity
intracluster, 64, 68
minimum, maximum, or average, 64
Similarity based criterion functions, 90–95
intracluster errors, 90
intracluster similarity, 91
pairwise distance, 90
sum of centroid similarity, 91
sum of squared errors (SSE), 90
Similarity based feature selection, 122
Similarity between cluster centroids, 64
Similarity search, 36–42
bag-of-words approach, 41
cluster hypothesis, 36
cosine similarity, 24, 25, 36–38
document resemblance, 41–42
jaccard coefficient, 38–39
jaccard metric, 39
jaccard similarity, 38–41
resemblance, 42
set-of-words approach, 41
shingles, 42
sketch, 42
Site visit, 167
Social networks analysis, 48–49
Software
WEKA, xiv
Clementine, xiv
Supervised learning framework, 115–116
Support, 198, 199
Support vector machine (SVM), 133

Target relations, 133
Term frequency (TF), 17
Time stamp, creating, 159

User identification, 164–167
User session, 167

Variable extraction, 159
Vector space model, 21–23
Visit actions, 177
Visit time, 179

Web content mining, 59–139
Web search, 1–46
Web server log files (web logs), 148–151
Web structure mining, 1–58
Web usage mining, 141–212
definition of, 143–144
exploratory data analysis (EDA), 177–190. See also Exploratory data analysis
introduction to, 143–155
auxiliary information, 154
clickstream analysis, 147–148
common log format, 151
authuser field, 151
identification field, 151
cross-industry standard process for data mining (CRISP-DM), 144–147.
See also Cross-industry standard process for data mining
extended common log format, 151–153
referrer field, 152
user agent field, 152
web log record, example of, 153
web server log files, 148–151
date/time field, 149
HTTP request field, 149
remote host field, 149
status code field, 150
transfer volume (bytes) field, 151
web log file, 148
Microsoft IIS log format, 153–154
modeling for, 191–212. See also Modeling for web usage mining
preprocessing for, 156–176. See also Preprocessing for web usage mining
WEKA, xiv
White box approach, xii
Within-cluster variation, 194
Z-score standardization, 194