

classification

Christopher D. Manning

search

Prabhakar Raghavan

Hinrich Schütze

precision

crawler

links

spam

# Introduction to Information Retrieval

recall

query

clustering

svm

025.04 MAN/I

CASMTVK

Books



2 7 7

index

web

xml

language model

CAMBRIDGE

ranking





**CAMBRIDGE**  
UNIVERSITY PRESS

4381/4 Ansari Road, Daryaganj, Delhi 110002, India

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

[www.cambridge.org](http://www.cambridge.org)

Information on this title: [www.cambridge.org/9780521865715](http://www.cambridge.org/9780521865715)

© Cambridge University Press 2008

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2008

First South Asian edition 2012

Reprinted 2013, 2014

This South Asian edition is based on Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze / *Introduction to Information Retrieval* / 9780521865715 / 2008

Printed in India by Shree Maitrey Printech Pvt. Ltd., Noida

*Library of Congress Cataloging in Publication data*

Manning, Christopher D.

*Introduction to information retrieval* / Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze.

p. cm.

Includes bibliographical references and index.

ISBN 978-0-521-86571-5 (hardback)

1. Text processing (Computer science) 2. Information retrieval. 3. Document clustering. 4. Semantic Web. I. Raghavan, Prabhakar. II. Schütze, Hinrich.

III. Title.

QA76.9.T48M26 2008

025.04 – dc22 2008001257

ISBN-13 978-1-107-66639-9 (paperback)

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party Internet Web sites referred to in this publication and does not guarantee that any content on such Web sites is, or will remain, accurate or appropriate. Information regarding prices, travel timetables, and other factual information given in this work are correct at the time of first printing, but Cambridge University Press does not guarantee the accuracy of such information thereafter.

## Contents

<i>Table of Notation</i>	page xi
<i>Preface</i>	xv
<b>1 Boolean retrieval</b>	1
1.1 An example information retrieval problem	3
1.2 A first take at building an inverted index	6
1.3 Processing Boolean queries	9
1.4 The extended Boolean model versus ranked retrieval	13
1.5 References and further reading	16
<b>2 The term vocabulary and postings lists</b>	18
2.1 Document delineation and character sequence decoding	18
2.2 Determining the vocabulary of terms	21
2.3 Faster postings list intersection via skip pointers	33
2.4 Positional postings and phrase queries	36
2.5 References and further reading	43
<b>3 Dictionaries and tolerant retrieval</b>	45
3.1 Search structures for dictionaries	45
3.2 Wildcard queries	48
3.3 Spelling correction	52
3.4 Phonetic correction	58
3.5 References and further reading	59
<b>4 Index construction</b>	61
4.1 Hardware basics	62
4.2 Blocked sort-based indexing	63
4.3 Single-pass in-memory indexing	66
4.4 Distributed indexing	68
4.5 Dynamic indexing	71



4.6 Other types of indexes	73
4.7 References and further reading	76
<b>5 Index compression</b>	<b>78</b>
5.1 Statistical properties of terms in information retrieval	79
5.2 Dictionary compression	82
5.3 Postings file compression	87
5.4 References and further reading	97
<b>6 Scoring, term weighting, and the vector space model</b>	<b>100</b>
6.1 Parametric and zone indexes	101
6.2 Term frequency and weighting	107
6.3 The vector space model for scoring	110
6.4 Variant tf-idf functions	116
6.5 References and further reading	122
<b>7 Computing scores in a complete search system</b>	<b>124</b>
7.1 Efficient scoring and ranking	124
7.2 Components of an information retrieval system	132
7.3 Vector space scoring and query operator interaction	136
7.4 References and further reading	137
<b>8 Evaluation in information retrieval</b>	<b>139</b>
8.1 Information retrieval system evaluation	140
8.2 Standard test collections	141
8.3 Evaluation of unranked retrieval sets	142
8.4 Evaluation of ranked retrieval results	145
8.5 Assessing relevance	151
8.6 A broader perspective: System quality and user utility	154
8.7 Results snippets	157
8.8 References and further reading	159
<b>9 Relevance feedback and query expansion</b>	<b>162</b>
9.1 Relevance feedback and pseudo relevance feedback	163
9.2 Global methods for query reformulation	173
9.3 References and further reading	177
<b>10 XML retrieval</b>	<b>178</b>
10.1 Basic XML concepts	180
10.2 Challenges in XML retrieval	183
10.3 A vector space model for XML retrieval	188
10.4 Evaluation of XML retrieval	192

10.5 Text-centric versus data-centric XML retrieval	196
10.6 References and further reading	198
<b>11 Probabilistic information retrieval</b>	<b>201</b>
11.1 Review of basic probability theory	202
11.2 The probability ranking principle	203
11.3 The binary independence model	204
11.4 An appraisal and some extensions	212
11.5 References and further reading	216
<b>12 Language models for information retrieval</b>	<b>218</b>
12.1 Language models	218
12.2 The query likelihood model	223
12.3 Language modeling versus other approaches in information retrieval	229
12.4 Extended language modeling approaches	230
12.5 References and further reading	232
<b>13 Text classification and Naive Bayes</b>	<b>234</b>
13.1 The text classification problem	237
13.2 Naive Bayes text classification	238
13.3 The Bernoulli model	243
13.4 Properties of Naive Bayes	245
13.5 Feature selection	251
13.6 Evaluation of text classification	258
13.7 References and further reading	264
<b>14 Vector space classification</b>	<b>266</b>
14.1 Document representations and measures of relatedness in vector spaces	267
14.2 Rocchio classification	269
14.3 $k$ nearest neighbor	273
14.4 Linear versus nonlinear classifiers	277
14.5 Classification with more than two classes	281
14.6 The bias-variance tradeoff	284
14.7 References and further reading	291
<b>15 Support vector machines and machine learning on documents</b>	<b>293</b>
15.1 Support vector machines: The linearly separable case	294
15.2 Extensions to the support vector machine model	300
15.3 Issues in the classification of text documents	307
15.4 Machine-learning methods in ad hoc information retrieval	314
15.5 References and further reading	318



viii	Contents	Contents
<b>16 Flat clustering</b>		321
16.1 Clustering in information retrieval		322
16.2 Problem statement		326
16.3 Evaluation of clustering		327
16.4 K-means		331
16.5 Model-based clustering		338
16.6 References and further reading		343
<b>17 Hierarchical clustering</b>		346
17.1 Hierarchical agglomerative clustering		347
17.2 Single-link and complete-link clustering		350
17.3 Group-average agglomerative clustering		356
17.4 Centroid clustering		358
17.5 Optimality of hierarchical agglomerative clustering		360
17.6 Divisive clustering		362
17.7 Cluster labeling		363
17.8 Implementation notes		365
17.9 References and further reading		367
<b>18 Matrix decompositions and latent semantic indexing</b>		369
18.1 Linear algebra review		369
18.2 Term–document matrices and singular value decompositions		373
18.3 Low-rank approximations		376
18.4 Latent semantic indexing		378
18.5 References and further reading		383
<b>19 Web search basics</b>		385
19.1 Background and history		385
19.2 Web characteristics		387
19.3 Advertising as the economic model		392
19.4 The search user experience		395
19.5 Index size and estimation		396
19.6 Near-duplicates and shingling		400
19.7 References and further reading		404
<b>20 Web crawling and indexes</b>		405
20.1 Overview		405
20.2 Crawling		406
20.3 Distributing indexes		415
20.4 Connectivity servers		416
20.5 References and further reading		419

Contents	ix
<b>21 Link analysis</b>	421
21.1 The Web as a graph	422
21.2 PageRank	424
21.3 Hubs and authorities	433
21.4 References and further reading	439
<b>Bibliography</b>	441
<b>Index</b>	469

Symbol	Page	Meaning
$\alpha$	40	Learning rate
$\beta$	20	Closest neighbor function
$\gamma$	224	Supervised learning model
$\delta$	371	Aggregation
$\epsilon$	24	Control of a class (or logistic) in a linear model
$\eta$	108	Training example
$\theta$	174	Singular value
$\lambda$	10	$\lambda$ -norm on the complexity of a function
$\mu$	121	Cluster in clustering
$\nu$	321	Clustering in neural networks
$\rho$	161	The value of $\gamma$ for which $f$ reaches its maximum
$\sigma$	164	The value of $\gamma$ for which $f$ reaches its minimum
$\tau$	217	Class or category classification
$\omega$	52	The collection frequency of term $t$ (the total number of times the term appears in the document collection)
$\phi$	237	Set $\{c_1, \dots, c_n\}$ of all classes
$\psi$	216	A random variable that takes as values elements of $C$
$\chi$	200	Term-document matrix
$\xi$	1	Index of the $i$ th document in the collection $D$
$\zeta$	25	A document
$\eta$	100	Document vector query vector
$\theta$	326	Set $\{d_1, \dots, d_n\}$ of all documents
$\iota$	200	Set of documents that is a subset
$\kappa$	217	Set $\{d_1, \dots, d_n\}$ of all documents in Chapters 13–15



- Zhang, Jiangong, Xiaohui Long, and Torsten Suel. 2007. Performance of compressed inverted list caching in search engines. In *Proc. CIKM*. ACM Press. [98]
- Zhang, Tong, and Frank J. Oles. 2001. Text categorization based on regularized linear classification methods. *IR* 4(1):5–31. URL: [citeseer.ist.psu.edu/zhang00text.html](http://citeseer.ist.psu.edu/zhang00text.html). [319]
- Zhao, Ying, and George Karypis. 2002. Evaluation of hierarchical clustering algorithms for document datasets. In *Proc. CIKM*, pp. 515–524. ACM Press. DOI: <http://doi.acm.org/10.1145/584792.584877>. [367]
- Zipf, George Kingsley. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley. [97]
- Zobel, Justin. 1998. How reliable are the results of large-scale information retrieval experiments? In *Proc. SIGIR*, pp. 307–314. [160]
- Zobel, Justin, and Philip Dart. 1995. Finding approximate matches in large lexicons. *Software Practice and Experience* 25(3):331–345. URL: [citeseer.lfi.unizh.ch/zobel95finding.html](http://citeseer.lfi.unizh.ch/zobel95finding.html). [60]
- Zobel, Justin, and Philip Dart. 1996. Phonetic string matching: Lessons from information retrieval. In *Proc. SIGIR*, pp. 166–173. ACM Press. [60]
- Zobel, Justin, and Alistair Moffat. 2006. Inverted files for text search engines. *ACM Computing Surveys* 38(2). [17, 76, 98, 122]
- Zobel, Justin, Alistair Moffat, Ross Wilkinson, and Ron Sacks-Davis. 1995. Efficient retrieval of partial documents. *IP&M* 31(3):361–377. DOI: [http://dx.doi.org/10.1016/0306-4573\(94\)00052-5](http://dx.doi.org/10.1016/0306-4573(94)00052-5). [199]
- Zukowski, Marcin, Sandor Heman, Niels Nes, and Peter Boncz. 2006. Super-scalar RAM-CPU cache compression. In *Proc. International Conference on Data Engineering*, p. 59. IEEE Computer Society. DOI: <http://dx.doi.org/10.1109/ICDE.2006.150>. [98]

## Index

- A/B test, 156
- Accents, 27–28
- Access control lists, 74
- Accumulator, 103, 115
- Accuracy, 143
- Active learning, 309
- Add-one smoothing, 240
- Ad hoc retrieval
- defined, 4–5
  - evaluation of, 139–141
  - machine learning methods, 314–318, 320
- Adjacency tables, 417
- Adjusted Rand index, 330
- Adversarial information retrieval, 392
- Akaike information criterion (AIC), 337
- Algebra, linear, review, 369–373
- Algorithmic search, 393
- Anchor text, 389, 422–423
- Any-of classification, 238, 281
- Auxiliary index, 71
- Average-link clustering, 350, 356–358
- Back queues, 412–415
- Bag of words model. *See also* Unigram language model
- defined, 107, 113–114
- Balanced F measure, 144. *See also* F measure
- Bayes error rate, 277
- Bayesian networks, 215–216
- Bayesian prior, 208, 210
- Bayesian smoothing, 226
- Bayes Optimal Decision Rule, 203
- Bayes risk, 203
- Bayes' Rule, 202
- Bernoulli model, 243–251
- Best-merge persistence, 355
- Bias, 286
- Bias-variance tradeoff, 284–289, 292
- Biclustering, 345
- Bigram language model, 221–222
- Binary Independence Model (BIM), 204–212, 229–230
- Binary search tree, 46, 47
- Biword indexes, 36–38
- Blind relevance feedback, 171–172
- Blocked sort-based indexing algorithm (BSBI), 63–66, 75
- Blocked storage described, 85–87
- Blogs, 178
- BM25 weights, 213–215
- Boolean retrieval
- defined, xvi
  - model, 4
  - principles, 3–6
  - query processing, 9–13
  - ranked retrieval *vs.*, 13–16
  - tokenization, 26
  - vector space model interactions, 136
- Boosting, 264
- Bottom-up clustering. *See* hierarchical agglomerative clustering (HAC)
- Bowtie structure, WWW, 389
- Break-even point, 148, 261, 306
- BSBI (blocked sort-based indexing algorithm), 66, 75
- B-trees, 47–48
- Buckshot algorithm, 366
- Buffer, 62
- Caching
- compression and, 78
  - defined, 62
  - in search systems, 135, 409, 411
  - variable length arrays and, 9
- Capitalization, 26
- Capture-recapture method, 396–400



- Turpin, Andrew, and William R. Hersh. 2001. Why batch and user evaluations do not give the same results. In *Proc. SIGIR*, pp. 225–231.
- Turpin, Andrew, and William R. Hersh. 2002. User interface effects in past batch versus user experiments. In *Proc. SIGIR*, pp. 431–432.
- Turpin, Andrew, Yohannes Tsegay, David Hawking, and Hugh E. Williams. 2007. Fast generation of result snippets in web search. In *Proc. SIGIR*, pp. 127–134. ACM Press. [161]
- Turtle, Howard. 1994. Natural language vs. Boolean query evaluation: A comparison of retrieval performance. In *Proc. SIGIR*, pp. 212–220. ACM Press. [15]
- Turtle, Howard, and W. Bruce Croft. 1989. Inference networks for document retrieval. In *Proc. SIGIR*, pp. 1–24. ACM Press. [215]
- Turtle, Howard, and W. Bruce Croft. 1991. Evaluation of an inference network-based retrieval model. *TOIS* 9(3):187–222. [215]
- Turtle, Howard, and James Flood. 1995. Query evaluation: strategies and optimizations. *IP&M* 31(6):831–850. doi: [http://dx.doi.org/10.1016/0306-4573\(95\)00020-H](http://dx.doi.org/10.1016/0306-4573(95)00020-H). [123]
- Vaithyanathan, Shivakumar, and Byron Dom. 2000. Model-based hierarchical clustering. In *Proc. UAI*, pp. 599–608. Morgan Kaufmann. [368]
- van Rijsbergen, C. J. 1979. *Information Retrieval*, 2nd edition. Butterworths. [159, 198, 203, 213, 216]
- van Rijsbergen, C. J. 1989. Towards an information logic. In *SIGIR*, pp. 77–86. ACM Press. doi: <http://doi.acm.org/10.1145/75334.75344>. [xviii]
- van Zwol, Roelof, Jeroen Baas, Herre van Oostendorp, and Frans Wiering. 2006. Bricks: The building blocks to tackle query formulation in structured document retrieval. In *Proc. ECIR*, pp. 314–325. [200]
- Vapnik, Vladimir N. 1998. *Statistical Learning Theory*. Wiley-Interscience. [319]
- Vittaut, Jean-Noël, and Patrick Gallinari. 2006. Machine learning ranking for structured information retrieval. In *Proc. ECIR*, pp. 338–349. [199]
- Voorhees, Ellen M. 1985a. The cluster hypothesis revisited. In *Proc. SIGIR*, pp. 188–196. ACM Press. [344]
- Voorhees, Ellen M. 1985b. *The effectiveness and efficiency of agglomerative hierarchic clustering in document retrieval*. Technical Report TR 85-705, Cornell. [367]
- Voorhees, Ellen M. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *IP&M* 36:697–716. [160]
- Voorhees, Ellen M., and Donna Harman (eds.). 2005. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press. [159, 453, 461]
- Wagner, Robert A., and Michael J. Fischer. 1974. The string-to-string correction problem. *JACM* 21(1):168–173. doi: <http://doi.acm.org/10.1145/321796.321811>. [59]
- Ward Jr., J. H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58:236–244. [367]
- Wei, Xing, and W. Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proc. SIGIR*, pp. 178–185. ACM Press. doi: <http://doi.acm.org/10.1145/1148170.1148204>. [384]
- Weigend, Andreas S., Erik D. Wiener, and Jan O. Pedersen. 1999. Exploiting hierarchy in text categorization. *IR* 1(3):193–216. [319]
- Weston, Jason, and Chris Watkins. 1999. Support vector machines for multi-class pattern recognition. In *Proc. European Symposium on Artificial Neural Networks*, pp. 219–224. [319]
- Williams, Hugh E., and Justin Zobel. 2005. Searchable words on the web. *International Journal on Digital Libraries* 5(2):99–105. doi: <http://dx.doi.org/10.1007/s00799-003-0050-z>. [97]
- Williams, Hugh E., Justin Zobel, and Dirk Bahle. 2004. Fast phrase querying with combined indexes. *TOIS* 22(4):573–594. [41]
- Witten, Ian H., and Timothy C. Bell. 1990. Source models for natural language text. *International Journal Man-Machine Studies* 32(5):545–579. [97]

- Witten, Ian H., and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edition. Morgan Kaufmann. [342]
- Witten, Ian H., Alistair Moffat, and Timothy C. Bell. 1999. *Managing Gigabytes: Compressing and Indexing Documents and Images*, 2nd edition. Morgan Kaufmann. [76, 97, 98]
- Wong, S. K. Michael, Yiyu Yao, and Peter Bollmann. 1988. Linear structure in information retrieval. In *Proc. SIGIR*, pp. 219–232. ACM Press. [320]
- Woodley, Alan, and Shlomo Geva. 2006. NLPX at INEX 2006. In *Proc. INEX*, pp. 302–311. [200]
- Xu, Jinxi, and W. Bruce Croft. 1996. Query expansion using local and global document analysis. In *Proc. SIGIR*, pp. 4–11. ACM Press. [177]
- Xu, Jinxi, and W. Bruce Croft. 1999. Cluster-based language models for distributed retrieval. In *Proc. SIGIR*, pp. 254–261. ACM Press. doi: <http://doi.acm.org/10.1145/312624.312687>. [344]
- Yang, Hui, and Jamie Callan. 2006. Near-duplicate detection by instance-level constrained clustering. In *Proc. SIGIR*, pp. 421–428. ACM Press. doi: <http://doi.acm.org/10.1145/1148170.1148243>. [344]
- Yang, Yiming. 1994. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In *Proc. SIGIR*, pp. 13–22. ACM Press. [291]
- Yang, Yiming. 1999. An evaluation of statistical approaches to text categorization. *IR* 1:69–90. [319]
- Yang, Yiming. 2001. A study of thresholding strategies for text categorization. In *Proc. SIGIR*, pp. 137–145. ACM Press. doi: <http://doi.acm.org/10.1145/383952.383975>. [292]
- Yang, Yiming, and Bryan Kisiel. 2003. Margin-based local regression for adaptive filtering. In *Proc. CIKM*, pp. 191–198. ACM Press. doi: <http://doi.acm.org/10.1145/956863.956902>. [292]
- Yang, Yiming, and Xin Liu. 1999. A re-examination of text categorization methods. In *Proc. SIGIR*, pp. 42–49. ACM Press. [265, 319]
- Yang, Yiming, and Jan Pedersen. 1997. Feature selection in statistical learning of text categorization. In *Proc. ICML*. [265]
- Yue, Yisong, Thomas Finley, Filip Radlinski, and Thorsten Joachims. 2007. A support vector method for optimizing average precision. In *Proc. SIGIR*. ACM Press. [320]
- Zamir, Oren, and Oren Etzioni. 1999. Grouper: A dynamic clustering interface to web search results. In *Proc. WWW*, pp. 1361–1374. Elsevier North-Holland. doi: [http://dx.doi.org/10.1016/S1389-1286\(99\)00054-7](http://dx.doi.org/10.1016/S1389-1286(99)00054-7). [344, 368]
- Zaragoza, Hugo, Djoerd Hiemstra, Michael Tipping, and Stephen Robertson. 2003. Bayesian extension to the language model for ad hoc information retrieval. In *Proc. SIGIR*, pp. 4–9. ACM Press. [232]
- Zavrel, Jakub, Peter Berck, and Willem Lavrijsen. 2000. Information extraction by text classification: Corpus mining for features. In *Proc. Workshop Information Extraction meets Corpus Linguistics*. URL: <http://www.cnts.ua.ac.be/Publications/2000/ZBL00>. Held in conjunction with LREC-2000. [292]
- Zha, Hongyuan, Xiaofeng He, Chris H. Q. Ding, Ming Gu, and Horst D. Simon. 2001. Bipartite graph partitioning and data clustering. In *Proc. CIKM*, pp. 25–32. ACM Press. [345, 368]
- Zhai, Chengxiang, and John Lafferty. 2001a. Model-based feedback in the language modeling approach to information retrieval. In *Proc. CIKM*. ACM Press. [231]
- Zhai, Chengxiang, and John Lafferty. 2001b. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. SIGIR*, pp. 334–342. ACM Press. [232]
- Zhai, Chengxiang, and John Lafferty. 2002. Two-stage language models for information retrieval. In *Proc. SIGIR*, pp. 49–56. ACM Press. doi: <http://doi.acm.org/10.1145/564376.564387>. [233]



- Cardinality in clustering, 327, 336–338
- Case-folding, 26
- CAS topics, 193
- Category, 237
- Centroid-based classification, 291
- Centroids
  - defined, 331
  - HAC, 350, 358–359
  - Rocchio classification, 269, 271
- Chaining in clustering, 352
- Chain rule, 202
- Champion lists, 127–128
- Character sequence decoding, 18–21
- $\chi^2$  feature selection, 256, 258
- Chinese, 23–24, 47
- Class boundary, 279
- Classes
  - defined, 238
  - maximum a posteriori, 239
- Classification. *See also* Text classification
  - any-of, 281
  - centroid-based, 291
  - defined, 234, 235
  - kNN (*See* k nearest neighbor classification (kNN))
  - multivalued, 281
  - one-of, 282
  - one-versus-all, 303
  - Rocchio (*See* Rocchio classification)
- Classification function, 237
- Classifiers
  - choosing, 308–309
  - defined, 237
  - performance, improving, 309–313
  - two-class, 259, 267, 292
- CLEF collection, 142
- Click spam, 394
- Clickstream mining, 172
- Clickthrough log analysis, 156, 172
- Cliques, 351
- Cloaking, in spamming, 391
- Cluster-based classification, 291
- Cluster hypothesis, 322–323, 325, 344
- Clustering
  - average-link, 350, 358
  - cardinality in, 327, 338
  - centroid-based, 362, 367
  - chaining in, 352
  - complete-link HAC, 360
  - divisive, 362–363
  - exclusive vs. exhaustive, 327
  - flat (*See* Flat clustering)
  - function notations, xi
  - group-average agglomerative, 350, 358, 360, 362, 367
  - hard, 322
  - hierarchical, 346 (*See also* Hierarchical clustering)
  - minimum variance, 367
  - model-based, 338–342
  - optimal, 362
  - overview, 322–326
  - single-link HAC, 359, 360, 362
  - spectral, 368
  - top-down, 363
- Clusters
  - defined, 68, 321
  - labeling, 363–365, 367–368
  - pruning, 129–131
- Co-clustering, 345
- Collections
  - clustering, 325
  - defined, 4
  - frequency, 25, 108–109
  - residual defined, 171
  - statistics, large, 75
- Combination schemes, 40–42
- Combination similarity, 347, 351, 360
- Complete-linkage clustering. *See* Complete-link clustering
- Component coverage, 193–194
- Compound nouns, 24
- Compound-splitter, 24
- Compression
  - of dictionaries, 82–87, 102
  - of docIDs, 88
  - lossless/lossy, 80
  - parameter-free, 92
  - parameterized, 98
  - of postings list, 87–95
- Compression/indexes
  - Heaps' law, 80–82, 276–277
  - overview, 78
  - Zipf's law, 82–83
- Concept drift, 249
- Conditional independence assumption, 246
- Confusion matrix, 283
- Connected components, 351
- Connectivity queries, 416
- Connectivity servers, 419
- Content management systems, 77
- Content seen module, 410–411
- Context, XML, 181
- Context resemblance, 190
- Contiguity hypothesis, 266
- Continuation bit, 89
- Corpus, 4
- Cosine similarity, 111, 112, 121, 344

- CO topics, 193
- CPC (cost per click), 393
- CPM (cost per mil), 393
- Cranfield collection, 141–142
- Cross-entropy, 232
- Cross-language information retrieval, 142, 384
- Cumulative gain, 149
- Databases
  - communication with, 77
  - relational, 178–179, 197
- $\Delta$ -codes, 96, 98
- Decision boundaries
  - defined, 269
  - kNN, 274
- Decision hyperplanes, 267, 278
- Decision trees, 261
- Dendrograms
  - complete-link clustering, 352
  - described, 347, 348
- Development sets, 262
- Development test collection, 141
- Diacritics, 28
- Dice coefficient, 150
- Dictionaries
  - compression of, 87, 102
  - in inverted indexes, 5–7
  - search structures for, 45–47
- Differential cluster labeling, 365
- Digital libraries, 178
- Discrete-time stochastic processes, 425
- Disk seek, 62
- Distortion, 336
- Distributed crawling, 419
- Distributed index, 68
- Distributed indexing, 67–70, 415–416
- Distributed information retrieval, 70, 416
- Divisive clustering, 363
- DNS resolution, 411–412
- DNS resolution module, 408
- DNS server, 411
- DocIDs
  - compression of, 88
  - in inverted indexes, 7
  - in postings list intersection operations, 10
- Document-at-a-time scoring, 129
- Document collection. *See* Collections
- Document likelihood model, 231
- Document-partitioned index, 68
- Documents
  - character sequence decoding, 21
  - classification of (*See* Text classification)
  - defined, 4
  - delineation of, 21
  - frequency defined, 7
  - function notations, xi
  - partitioning, 416
  - relevant, retrieving, xvii
  - unit, choosing, 20–21
  - vector, defined, 109–110
- Document space, 237
- Document zones, 312–313
- Doorway pages, 392
- Dot products
  - described, 110–113
  - in SVMs, 298
- Duplicate elimination modules, 408
- Dynamic indexing, 71
- Dynamic summary, 157
- East Asian languages, 43. *See also* Chinese; Japanese
- Edit distance, 53–55
- Effectiveness
  - assessment of, 5
  - text classification, 259, 261
- Efficiency, 259
- Eigen decomposition, 372
- Eigenvalues, 370, 425
- 11-point interpolated average precision, 146
- Email
  - document units, 20
  - sorting, 2, 235
- EM algorithm, 339–341
- Enterprise resource planning, 77
- Enterprise search, 61
- Entropy, 91, 330
- Equivalence classes, 26
- Ergodic Markov Chain, 427
- Euclidean distance, 121, 296–297, 344
- Euclidean length, 111
- Evaluation of retrieval systems
  - A/B test, 156
  - ad hoc, 141
  - clustering, 327–331
  - F measure, 144, 331
  - interpolated precision, 145
  - kappa statistic, 151, 152, 160
  - keyword-in-context snippets, 158
  - MAP, 147
  - marginal relevance, 154
  - normalized discounted cumulative gain, 149
  - overview, 141
  - pooling, 151
  - precision at k, 148
  - precision-recall curve, 145, 146



Evaluation of retrieval systems (*cont.*)  
 probabilistic information retrieval,  
 212–213  
 ranked sets, 145–151  
 relevance assessment, 154  
 relevance feedback, 170–171  
 results snippets, 159  
 ROC curve, 149  
 R-precision, 148, 160  
 sensitivity, 149  
 specificity, 149  
 summarization, static *vs.* dynamic, 157  
 system quality/user utility, 156  
 test collections, standard, 142  
 text classification, 258–263  
 text summarization, 157  
 unranked sets, 142–145  
 XML retrieval, 192–196  
 Evidence accumulation, 134  
 Exclusive clustering, 327  
 Exhaustive clustering, 327  
 Expectation-Maximization (EM)  
 algorithm, 340, 341  
 Expectation step, 340  
 Expected edge density, 344–345  
 Extended query, 187–188  
 Extensible Markup Language. *See* XML  
 External criterion of quality, 328, 329  
 External sorting algorithm, 63  
 False negative, 330  
 False positive, 330  
 Feature engineering, 311  
 Feature selection/text classification  
 $\chi^2$ , 255–256  
 frequency-based, 257  
 greedy, 258  
 method comparison, 257–258  
 multiple classifiers, 257  
 mutual information, 252–255  
 noise feature, 251  
 overfitting, 251  
 overview, 251–252  
 in performance improvement,  
 310–312  
 statistical significance, 256  
 Fetch modules, 408  
 Field, 101  
 Filtering, 234, 291  
 First story detection, 362  
 Flat clustering  
 Akaike information criterion, 337  
 cardinality in, 327, 338  
 classification *vs.*, 321

collections, 325  
 defined, 321  
 distortion, 336  
 evaluation of, 331  
 exhaustive, 327  
 Expectation-Maximization algorithm,  
 340, 341  
 expectation step, 340  
 external criterion of quality, 328, 329  
 HAC *vs.*, 367  
 internal criterion of quality, 327  
 K means, 331–338  
 K-medoids, 336  
 in language models, 325  
 maximization step, 340  
 model complexity, 336  
 normalized mutual information, 329  
 objective functions, 326  
 outliers, 334  
 partitional, 326–327  
 purity, 328, 329  
 Rand index, adjusted, 330  
 residual sum of squares, 337  
 scatter-gather, 323, 324, 344  
 search result, 323  
 seeds, 332  
 singleton, 334  
 soft, 322, 382  
 unsupervised learning, 321  
 F measure, 144, 160, 331  
 Focused retrieval, 199  
 Free text, 100, 136–137  
 Free text query  
 parsing functions, designing, 133–134  
 tokenization, 26  
 in vector retrieval models, 13  
 Frequency-based feature selection, 257  
 Frobenius norm, 376  
 Front coding, 86, 87  
 Front queues, 415  
 Functional margins, 296  
 GAAC. *See* Group-average  
 agglomerative clustering  
 $\gamma$  encoding, 90–95  
 Gaps, encoding, 88  
 Generative model, 218–220  
 Geometric margin, 297  
 Global champion list, 128  
 Gold standard, 140  
 Golomb codes, 98  
 GOV2 collection, 142  
 Greedy feature selection, 258  
 Grepping, 3

Ground truth, 140  
 Group-average agglomerative  
 clustering, 350, 358, 360, 362, 367  
 Group-average clustering. *See*  
 Group-average agglomerative  
 clustering  
 HAC. *See* hierarchical agglomerative  
 clustering (HAC)  
 Hard assignment, 322  
 Hard clustering, 326  
 Harmonic numbers, 93  
 Hashing, 46, 86–87  
 Heaps' law, 82, 277  
 Held-out data, 262  
 Hierarchical agglomerative clustering  
 (HAC)  
 algorithm comparison, 362  
 best-merge persistence, 355  
 Buckshot algorithm, 366  
 centroids, 350, 359, 362, 367  
 chaining in, 352  
 cliques, 351  
 cluster-internal labeling, 365  
 combination similarity, 347, 360  
 complete-link clustering, 359, 360, 362,  
 367  
 connected components, 351  
 dendrograms, 347, 348, 352  
 differential cluster labeling, 365  
 divisive, 363  
 first story detection, 362  
 flat *vs.*, 367  
 group-average, 350, 358, 360, 362,  
 367  
 inversions, 347, 359  
 monotonicity, 347  
 next-best merge (NBM) arrays, 355  
 novelty detection, 362  
 optimality, 362  
 outliers, 353  
 overview, 347–349  
 priority queue algorithm, 353, 354  
 single-link clustering, 359–360, 362  
 time complexity, 353–356  
 top-down, 363  
 Hierarchical classification, 319  
 Hierarchical clustering, 346  
 agglomerative (*See* hierarchical  
 agglomerative clustering (HAC))  
 applications, 346–347  
 defined, 321  
 probabilistic interpretation of, 368  
 R environment support for, 368  
 Hierarchical Dirichlet processes, 384  
 Hierarchy, 346  
 Highlighting, 186  
 HITS (hyperlink-induced topic search),  
 435, 437, 439  
 Host splitters, 410  
 HTML, 385  
 http, 385  
 Hub score, 433–439  
 Hyperlink-induced topic search (HITS),  
 435, 437, 439  
 Hyperlinks, 389. *See also* Link analysis  
 Hyphenation and tokenization, 24  
 Ide dec-hi, 167  
 IDF. *See* Inverse document frequency  
 (IDF)  
 IID. *See* Independent and identically  
 distributed (IID)  
 Images, searching for. *See* Relevance  
 feedback  
 Impact ordering, 129  
 Implicit relevance feedback, 172  
 Incidence matrix, 374  
 Independence, 255  
 Independent and identically distributed  
 (IID), 262  
 Index construction  
 BSBI, 66, 75  
 distributed indexes, 68, 419  
 resources, 76  
 Indexer, 61  
 Indexes  
 biword, 38  
 defined, 3  
 document-partitioned, 70, 416  
 k-gram, 50–51, 55–57, 311  
 next word, 41  
 parametric, 101–107  
 permuterm, 49–50  
 positional, 38–40  
 size/estimation, 400  
 term-partitioned, 70  
 zone, 107  
 Indexing  
 defined, 61  
 distributed, 70, 416  
 granularity, 20  
 latent semantic, 378–382  
 unit defined, 184  
 INEX, 196  
 Informational queries, 395  
 Information gain, 264  
 Information need, 5, 140



- Information retrieval
  - hardware issues, 63
  - history of, 17
  - overview, 3, xvi
  - search system components, 135, 135
  - terms, statistical properties of, 82
- In-links, 389
- Inner product. *See* Dot products
- Instance-based learning, 276
- Internal criterion of quality, 327
- Interpolated precision, 145
- Intersection, postings list, 10, 36
- Inter-similarity, 350
- Inverse document frequency (IDF), 109, 190, 209
- Inversions
  - defined, 64
  - in HAC, 347, 358
- Inverted file. *See* Inverted index; Postings list
- Inverted index
  - Boolean query processing, 13
  - building principles, 9
  - described, 6
  - $\gamma$  encoding, 90, 95
  - kNN classification in, 277
- Inverter, 69–70
- IP address, 411
- Jaccard coefficient, 56, 401
- Japanese, 29–30
- Journal influence weight, 439
- Kappa statistic, 151, 152, 160
- Kernel function, 305
- Kernels
  - Mercer, 305
  - polynomial, 305
  - quadratic, 305
  - radial basis functions, 305
- Kernel trick, 304
- Keys, 46
- Key-value pairs, 68
- Keyword-in-context (KWIC) snippets, 158
- k-gram index
  - described, 51
  - spelling correction in, 57
  - word matching in, 311
- K means, 338
- K-medoids, 336
- k nearest neighbor classification (kNN)
  - algorithm, 273–275
  - Bayes error rate, 277
  - bias in, 286–287
  - decision boundaries, 274
  - described, 267, 291–292
  - effectiveness, 261, 292
  - instance-based learning, 276
  - memory-based learning, 276
  - memory capacity, 287
  - multinomial Naive Bayes *vs.*, 249
  - as nonlinear classification, 280–281
  - testing/training capacity, 302
  - time complexity/optimal, 275–277
  - variance, 287
  - Voronoi tessellation, 273, 274
- KNN classification. *See* K nearest neighbor classification (kNN)
- Kruskal's algorithm, 367
- Kullback-Leibler divergence, 231, 344
- KWIC (keyword-in-context), 158
- Labeling
  - of clusters, 368
  - defined, 236
- Language, of an automaton, 219
- Language identification, 22
- Language issues, relevance feedback, 169–170
- Language models
  - Bayesian smoothing, 226
  - BIM/XML *vs.*, 230
  - clustering in, 325
  - defined, 219, 224
  - distributions, multinomial, 222–223
  - document likelihood, 231
  - extended approaches, 230–232
  - finite automata and, 220
  - Kullback-Leibler divergence, 231
  - likelihood ratio, 220
  - linear interpolation, 226
  - overview, 218
  - query likelihood, 223–229
  - tf-idf weighting *vs.*, 228
  - translation, 232
  - types of, 222
- Laplace smoothing, 240
- Latent Dirichlet Allocation (LDA), 384
- Latent semantic analysis (LSA), 379
- Latent semantic indexing (LSI), 382
- LDA (Latent Dirichlet Allocation), 384
- L2 distance, 121, 297, 344
- Learning algorithm described, 103–106. *See also* Weighted zone scoring
- Learning error, 285
- Learning method, 237
- Lemma, 31
- Lemmatization described, 30–33
- Lemmatizer, 32

- Length-normalization, 111
- Levenshtein distance, 55
- Lexicalized subtrees, 188–189
- Lexicons in inverted indexes, 6
- Likelihood, 202
- Likelihood ratio, 220
- Linear algebra review, 373
- Linear classifiers, 267, 277–281
- Linear interpolation, 226
- Linear problem, 279
- Linear separability, 280, 294–300
- Link analysis
  - anchor text, 389, 423
  - authority score, 439
  - ergodic Markov chain, 427
  - HITS, 435, 437
  - hub score, 439
  - Markov chains, 427
  - overview, 421
  - PageRank (*See* PageRank)
  - steady-state theorem, 427
- Link farms, 439
- Link spam, 421
- LLRUN, 98
- LM, 224. *See* Language models
- Logarithmic merging, 72
- Lossless compression, 80
- Lossy compression, 80
- Lovins stemmer, 32
- Low-rank approximation, 376–378
- LSA (latent semantic analysis), 379
- LSI (latent semantic indexing), 382
- Machine-learned relevance described, 106
- Machine learning methods, 318, 320
- Machine translation, 224
- Macroaveraging, 259–261
- MAP (mean average precision), 239
- Map phase, 69
- MapReduce, 69, 70, 76
- Marginal relevance, 154
- Marginal statistic, 152
- Margins, 295, 298
- Markov chains, 427
- Master node, 68
- Matrix decomposition
  - eigen, 372
  - eigenvalues, 370
  - Frobenius norm, 376
  - latent semantic indexing, 382
  - linear algebra review, 373
  - low-rank approximation, 378
  - reduced SVD, 374
  - singular value, 373–376
  - symmetric diagonal, 373, 374
  - theorems, 372–373
  - truncated SVD, 374
- Maximization step, 340
- Maximum a posteriori, 208
- Maximum likelihood estimate (MLE), 208, 224–227, 240, 252
- Mean average precision, 147
- Medoids, 336
- Memory-based learning, 276
- Memory capacity, 287
- Mercator crawler, 407, 419
- Mercer kernels, 305
- Merge algorithm, 10
- Merge postings list, 10, 65
- Metadata, 101
- Microaveraging, 261
- Minimum spanning tree, 367
- Minimum variance clustering, 367
- ModApte split, 259, 265
- Model-based clustering, 342
- Model complexity, 336
- Monotonicity, 347
- Multiclass classification, 282
- Multiclass SVMs, 303
- Multilabel classification, 281
- Multimodal class, 272
- Multinomial classification, 282
- Multinomial model, 242–243
- Multinomial Naive Bayes
  - Bernoulli model, 245, 251
  - bias in, 286
  - concept drift, 249
  - conditional independence
    - assumption, 246
    - as linear classifier, 278
    - optimal classifier, 250
  - positional independence assumption, 240, 247
  - properties, 251
  - in query likelihood models, 224
  - random variables X and U, 246
  - semi-supervised learning, 308
  - sparseness, 240
  - testing/training capacity, 302
  - in text classification, 243
  - variance, 287
- Multinomial NB. *See* Multinomial Naive Bayes
- Multivariate classification, 281
- Multivariate Bernoulli model, 245
- Mutual information, 255, 258



- Naive Bayes assumption, 168, 206
- Naive Bayes learning method, 237.  
*See also* Multinomial Naive Bayes;  
Multivariate Bernoulli model
- Named entity tagging, 178
- National Institute of Standards and  
Technology, 141
- Natural language processing  
issues in, 342  
lemmatizers in, 32  
text summarization, 313  
XML retrieval, 230
- Navigational queries, 395
- NDCG (normalized discounted  
cumulative gain), 149
- Near-duplicate search results, 400–403
- Nested elements, 185–186
- NEXI, 182
- Next-best merge (NBM) arrays, 355
- Next word index, 41
- N-gram language model, 43. *See also*  
Bigram language model; Unigram  
language model
- Nibble, 90
- NLP. *See* Natural language processing
- NMI. *See* Normalized mutual  
information (NMI)
- Noise documents, 279–280
- Noise feature, 251
- Nonlinear classifiers, 280, 303–306
- Nonlinear problem, 281
- Normalization  
in probability theory, 225  
term, 26–30  
tf weighting, 117  
URL, 409
- Normalized discounted cumulative gain  
(NDCG), 149
- Normalized mutual information (NMI),  
329
- Normalized tokens in inverted indexes,  
7
- Normal vectors, 270
- Notation, table of, xi
- Novelty detection, 362
- NTCIR collection, 142
- Objective function, 326, 332
- Odds, 203
- Odds ratio, 207
- Okapi BM25 weighting, 213
- 1/0 loss, 203
- One-of classification, 238, 263
- One-versus-all (OVA) classification,  
303

- Optimal classifier, 285
- Optimal clustering, 362
- Optimal learning method, 285
- Optimal weight, 106
- Ordering, 127–129
- Ordinal regression, 317
- Outliers, 334, 353
- Out-links, 389
- Overfitting, 287
- Overlap score measure, 109
- Oxford English Dictionary, 80
- PageRank  
computation, 427–430, 439  
described, 424–425  
ergodic Markov chain, 427  
Markov chains, 427  
personalized, 431  
principal left eigen vector, 425  
probability vectors, 426  
steady-state theorem, 427  
stochastic matrix, 425  
teleport operation, 424  
topic-specific, 430–432
- Paice stemmer, 32
- Paid inclusion, 391
- Parameter-free compression, 92
- Parameterized compression, 98
- Parameter tuning, 141, 291
- Parameter tying, 312
- Parametric indexes, 107
- Parametric search, 180
- Parser, 69
- Parsing functions, designing, 134
- Parsing modules, 408
- Partitional clustering, 327
- Partition rule, 202
- Passage retrieval, 199
- Patent databases, 178
- Performance, 259
- Permuterm index, 50
- Personalized PageRank, 431
- Pew Internet Survey 2004, xv
- Phonetic correction, 58–59
- Phrase index, 37
- Phrase queries, 36–42, 44, 137
- Phrase search, 14
- Pivoted document length normalization,  
121
- Pivot length, 120
- Pointwise mutual information, 265
- Polytomous classification, 282
- Polytopes, 274
- Pooling, 160
- Pornography filtering, 311

- Porter stemmer, 31, 32
- Positional independence assumption,  
240, 247
- Positional indexes, 40
- Posterior probability, 202
- Postfiltering, in k-gram indexes, 51
- Postings  
in block sort-based indexing, 64  
compression and, 79  
defined, 6, 79  
in inverted indexes, 7  
positional, 42
- Postings list  
compression of, 95  
described, 6  
intersection/merging, 10  
skip pointers, 36  
storage of, 9
- Power law, 389
- Precision, 5, 142
- Precision at k, 148
- Precision-recall curve, 145, 146
- Prefix-free code, 92
- Preprocessing, effects of, 80
- Principal direction divisive partitioning,  
368
- Priority queue algorithm, HAC, 353,  
354
- Prior probability, 202
- Probabilistic information retrieval  
Bayesian networks, 215  
Bayesian prior, 208  
Bayes Optimal Decision Rule, 203  
Binary Independence Model, 212  
evaluation, 213  
maximum a posteriori, 208  
maximum likelihood estimate, 208,  
227, 240, 252
- Naive Bayes assumption, 168, 206  
odds ratio, 207  
overview, 201  
probability theory principles,  
202–203  
pseudocounts, 208  
query generation, estimating, 227  
relative frequency, 208  
relevance feedback, 209–211  
Retrieval Status Value, 207  
tree-structured dependencies, 213
- Probability Ranking Principle, 204
- Probability vectors, 426
- Prototypes, 267
- Proximity operator, 14
- Proximity weighting, 132–133
- Pseudocounts, 208
- Pseudo-relevance feedback described,  
172
- Pull model, 291
- Purity, 328, 329
- Push model, 291
- Quadratic optimization, 298
- Queries. *See also* Terms  
BIM ranking function, deriving,  
205–207  
Boolean, 4, 13  
defined, 5  
expansion, 173–175  
extended, 188  
free text (*See* Free text query)  
generation probability, estimating, 227  
informational, 395  
navigational, 395  
optimization of, 10  
phrase (*See* Phrase queries)  
semistructured, 180  
simple conjunctive, 9  
structured, 180  
term highlighting, 159, 186  
transactional, 396  
user/web search, 395–396  
as vectors, 114
- Query-by-example, 183, 230
- Query likelihood model, 229
- Query parser, 134
- Query reformulation  
expansion, 175  
local vs. global, 162  
vocabulary tools for, 173
- Radial basis functions, 305
- Rand index, adjusted, 330, 344
- Random variables  
C, 248  
defined, 202  
U, 246  
X, 246
- Rank, of matrices, 369
- Ranked Boolean retrieval, 103. *See also*  
Weighted zone scoring
- Ranked retrieval models  
Boolean retrieval vs., 16  
described, 74  
evaluation of, 151
- Ranking/results  
BIM function, deriving, 207  
efficiency in, 124–125  
machine learning, 316–318
- Ranking SVM, 317
- Recall, 5, 143



- Reduced SVD, 378
- Reduce phase, 69
- Regression, 317
- Regular expressions, 3, 17
- Regularization, 301
- Relational databases, 179, 197
- Relative frequency, 208
- Relevance
  - assessment of, 154, 160
  - defined, 5
- Relevance feedback
  - applications, 170
  - evaluation of, 171
  - images, 163–164
  - implicit/indirect, 172
  - overview, 172–173
  - probabilistic models, 168, 211
  - pseudo-relevance, 172
  - Rocchio algorithm (*See* Rocchio algorithm)
  - text, 165
  - Web applications, 170
- R environment, 368
- Residual collection, 171
- Residual sum of squares (RSS), 332, 337
- Results snippets, 135
- Retrieval model, Boolean. *See* Boolean retrieval
- Retrieval Status Value, 207
- Reuters-21578 collection
  - confusion matrix, 283
  - described, 142
  - as linear, 279
  - text classification in, 259, 260, 261
- Reuters-RCV1 collection
  - blocked storage, 87
  - collection *vs.* document frequency, 109
  - construction of, 66, 75
  - described, 63–64, 77
  - dictionary-as-a-string storage, 83–85
  - dictionary compression, 95, 95
  - $\gamma$ -encoding, 92, 94
  - index compression, 95
  - preprocessing, effects of, 80
  - residual sum of squares, 337
  - Zipf's law, 82, 83
- RF. *See* Relevance feedback
- Robots Exclusion Protocol, 408
- Rocchio algorithm
  - applications, 170
  - overview, 163–168
- Rocchio classification
  - bias in, 286
  - centroids, 269–273
  - decision boundaries, 269
  - described, 273
  - effectiveness, 261, 292
  - as linear, 278
  - memory capacity, 287
  - multimodal class, 272
  - normal vectors, 270, 271
  - prototypes, 267
  - testing/training capacity, 302
  - variance, 287
- ROC curve, 149
- Routing, 234, 291
- R-precision, 148, 160
- RSS. *See* Residual sum of squares (RSS)
- Rule of 30, 79
- Rules in text classification, 236
- Scatter-Gather, 323, 324, 344
- Schema, 182
- Schema diversity/heterogeneity, 186–187
- Scoring
  - champion lists, 127, 128
  - cluster pruning, 131
  - document-at-a-time, 129
  - efficiency in, 125
  - functions, designing, 134
  - index elimination, 126–157
  - machine learning methods, 314–316
  - overview, 100
  - SimNoMerge, computing, 190, 191, 191
  - static quality scores, 129
  - top K document retrieval, 125–126
  - vector scores, computing, 114–116
  - vector space model (*See* Vector space model)
- Search advertising, 393, 394
- Search engines. *See also* Web index
  - components, 396
  - marketing, 394
  - optimizers, 392
- Search result clustering, 323
- Search results, 323
- Search system, complete, 135, 135. *See also* Web index
- Security, 74
- Seeds, 332
- Seed sets, 406
- Seek time, 62
- Segment file, 69
- Semistructured query, 180
- Semistructured retrieval, 179
- Semi-supervised learning, 308
- Sensitivity, 149

- Sentiment detection, 235
- Sequence model, 21–25, 28, 247
- Shingling, 403
- SimNoMerge, computing, 190, 191, 191
- Simple conjunctive queries, 9
- Single-label classification, 282
- Single-linkage clustering. *See* Single-link clustering
- Single-link clustering, 359, 360, 362
- Single-pass in-memory indexing (SPIMI), 67, 76
- Singleton cluster, 334, 347
- Singly-linked lists, 7
- Singular value decomposition (SVD), 376, 380
- Skip list, 36
- Skip pointers, 36
- Slack variables, 301
- SMART notation, 118
- Smoothing
  - add  $\alpha$ , 208
  - add-one, 240
  - add  $\frac{1}{2}$ , 208, 210, 211, 243
  - Bayesian, 226
  - Bayesian prior, 208, 210, 226
  - Laplace, 240
  - linear interpolation, 226
  - query generation estimation, 225–227
  - tf weighting, 117
- Snippet, 157
- Soft assignment, 322
- Soft clustering, 322, 382
- Soft margin classification, 300–303
- Sort-based multiway merge, 76
- Sorting, 7, 76
- Soundex algorithms, 59
- Spam
  - click, 394
  - filters, email, 2
  - link, 421
  - overview, 390–392
- Sparseness, 240
- Specificity, 149
- Spectral clustering, 368
- Speech recognition, 222
- Spelling correction, 51–58
- Spiders. *See* Web crawlers
- Spider traps, 405. *See also* Web crawlers
- SPIMI (single-pass in-memory indexing), 67, 76
- Splits, 68
- Sponsored search, 393
- Standing query, 234
- Static quality scores, 129
- Static summary, 157
- Static web pages, 388
- Statistical significance, 256
- Statistical text classification, 236
- Steady-state theorem, 427
- Stemming described, 33
- Stochastic matrix, 425
- Stop list, 25
- Stop words, 25–26
- Storage
  - blocked, 87
  - dictionary-as-a-string, 85
- Structural SVMs, 303
- Structural term, 189
- Structured document retrieval principle, 184
- Structured query, 180
- Structured retrieval, 178, 183–188
- Sublinear tf scaling, 117
- Summarization
  - in cluster labeling, 368
  - static *vs.* dynamic, 157
  - text, 157
- Supervised learning, 237
- Support vector, 294
- Support vector machines (SVMs)
  - active learning, 309
  - dot products in, 298
  - effectiveness, 262
  - Euclidean distance, 297
  - experimental results, 306–307
  - functional margins, 296
  - geometric margin, 297
  - kernel function, 305
  - kernels, polynomial, 305
  - kernel trick, 304
  - linear separability, 280, 300
  - margins, 294, 295
  - Mercer kernels, 305
  - multiclass, 303
  - nonlinear, 306
  - overview, 293
  - quadratic optimization, 298
  - radial basis functions, 305
  - ranking, 317
  - regularization, 301
  - slack variables, 301
  - soft margin classification, 303
  - structural, 303
  - testing/training capacity, 302
  - transductive, 309
  - weight vectors, 295
- SVD (singular value decomposition), 376, 380



- SVMs. *See* Support vector machines (SVMs)
- Symmetric diagonal decomposition, 373, 374
- Synonymy, 162
- Table of notation, xi
- Taxonomies, performance improvement, 310
- Teleport operation, 424
- Term-at-a-time, 115
- Term-document matrix  
defined, 4-5, 369  
singular value decomposition, 376, 380
- Term frequency  
benefits of, 15  
defined, 107  
weighting and, 107-110, 112
- TermID, 62
- Term normalization, 30
- Term-partitioned index, 70
- Terms. *See also* Queries  
BIM ranking function, deriving, 207  
defined, 3, 21  
function notations, xi  
partitioning, 416  
statistical properties of, 82  
tree-structured dependencies, 213  
vectors, weighting and, 113
- Term weighting. *See* Weighting
- Test data, 237
- Test set, 262
- Text, grepping, 3
- Text categorization. *See* Text classification
- Text classification  
Bernoulli model, 245, 251  
classes, 237, 238  
classifiers (*See* Classifiers; specific classifiers)  
decision trees, 261  
defined, 234  
development sets, 262  
document space, 237  
document zones, 313  
effectiveness, 259, 261  
email sorting (*See* Email)  
evaluation of, 263  
feature selection, 251-258  
held-out data, 262  
issues in, 307-313  
labeling, 236  
learning method, 237  
linear, 267, 281  
macroaveraging, 261  
microaveraging, 261  
ModApte split, 259  
multinomial Naive Bayes (*See* Multinomial Naive Bayes)  
nonlinear, 281  
overview, 237-238  
parameter tying, 312  
performance/efficiency, 259  
rules in, 236  
semi-supervised learning, 308  
sentiment detection, 235  
statistical, 236  
supervised learning, 237  
test sets, 237, 238  
training sets, 237, 238  
two-class classifier, 259, 267, 292  
vertical search engines, 235
- Text summarization, 157, 313
- TF. *See* Term frequency
- Tf-idf weighting, 116-121
- Thesauri  
automatic generation of, 175  
query expansion in, 175
- Tiered indexes described, 132-133
- Time complexity in HAC, 356
- Tokenization  
defined, 18  
hyphenation and, 24  
vocabulary/terms, determining, 25
- Tokens  
defined, 21  
in inverted indexes, 7  
normalization of, 30
- Top docs, 137
- Top-down clustering, 363  
classification of (*See* Text classification)  
standing queries *vs.*, 234  
in test collections, 142  
in XML retrieval, 193
- Topic-specific PageRank, 432
- Topic spotting. *See* Text classification
- Trailing wildcard query, 48
- Training set, 237, 238
- Transactional query, 396
- Transductive SVMs, 309
- Translation model, 232
- TREC collection, 142, 147
- Trec\_eval, 160
- Truecasing, 28
- Truncated SVD, 378, 381
- 20 Newsgroups, 142
- Two-class classifier, 259, 267, 292
- Type, 21

- Unary code, 90, 95
- Unigram language model. *See also* Bag of words model  
described, 222  
distributions, multinomial, 223  
multinomial Naive Bayes *vs.*, 243  
Union-find algorithm, 362, 403
- Universal code, 92
- Unsupervised learning, 321
- URLs  
defined, 386  
frontiers, 406, 407,  
normalization of, 409
- User document matrix, access control lists, 74
- Utility measure, 265
- Variable byte encoding, 88-90
- Variable length arrays, 9
- Variance, 287
- Vector space model. *See also* k nearest neighbor classification (kNN); Rocchio classification  
any-of classification, 281  
bias defined, 286  
bias-variance tradeoff, 289, 292  
class boundaries, 279  
confusion matrix, 283  
contiguity hypothesis, 266  
decision hyperplanes, 267, 278  
described, 110-116, 125  
document representation, 267-269  
learning error, 285  
linear classifiers, 267, 281  
linear separability, 280  
memory capacity, 287  
noise documents, 280  
nonlinear classifiers, 281  
one-of classification, 282  
optimal classifier, 250, 285  
optimal learning method, 285  
overfitting, 251, 287  
query operator interactions, 137  
relatedness measures, 269  
3+ classes, 281-283  
variance, 287  
XML retrieval, 188-192
- Vertical search engines, 235
- Vocabulary  
controlled, query expansion and, 175  
function notations, xi  
in inverted indexes, 6  
issues, relevance feedback, 170  
permuterm, 50
- Vocabulary/terms, determining  
common terms, dropping, 26  
lemmatization/stemming, 33  
normalization, 30  
tokenization, 25
- Voronoi tessellation, 273-274
- Ward's method, 367
- Web crawlers  
adjacency tables, 417  
back queues, 415  
connectivity servers, 419  
content seen module, 411  
distributed indexing, 70, 416  
distributing, 411  
DNS resolution, 412  
DNS resolution module, 408  
duplicate elimination modules, 408  
fetch modules, 408  
front queues, 415  
host splitters, 410  
Mercator, 407, 419  
operation/architecture, 406-410  
overview, 405-406  
parsing modules, 408  
Robots Exclusion Protocol, 408  
seed sets, 406  
URL frontiers, 406, 407, 415
- Web graphs, 389-390
- Web index. *See also* Search engines  
adversarial information retrieval, 392  
advertising/economic model, 392-394  
algorithmic search results, 393  
caching in, 135, 409, 411  
capture-recapture method, 400  
click spam, 394  
distributed indexing, 70, 416  
engine components, 396  
index size/estimation, 400  
informational queries, 395  
issues in, 2  
navigational queries, 395  
near-duplicate results, 403  
paid inclusion, 391  
query expansion, 175  
relevance feedback, 170  
search engine marketing, 394  
search engine optimizers, 392  
shingling, 403  
spam, 392  
sponsored, 393, 394  
transactional queries, 396  
user queries, 396



- Web pages  
 anchor text, 389  
 doorway, 392  
 dynamically generated, 388  
 hyperlinks, 389  
 power law, 82, 389  
 static, 388
- Weighted zone scoring  
 described, 102-104  
 learning algorithm, 106  
 optimal weight, 106
- Weighting  
 inverse document frequency, 109, 190, 209  
 Okapi BM25, 215  
 proximity, 133  
 SMART notation, 118  
 tf-idf (*See* Tf-idf weighting)
- Weight vectors, 295
- Westlaw, 14
- Wikipedia, 411
- Wildcard queries  
 defined, 45, 48  
 general, 48-50  
 k-gram index, 51  
 vector space model interactions, 136-137
- Within-point scatter, 343
- Word segmentation, 24
- World Wide Web. *See also under* Web  
 advertising/economic model, 394  
 background/history, 385-387  
 bowtie structure, 389, 390  
 characteristics, 387-392  
 HTML, 385  
 http, 385  
 paid inclusion, 391  
 spam, 392  
 URL, 386  
 web graphs, 390, 423
- XML, 19  
 attributes, 180  
 concepts, basic, 180-183  
 contexts, 181  
 data-centric, 179, 196-197  
 documents, decoding, 19  
 DOM, 181  
 DTD, 182  
 elements, 180  
 extended queries, 188  
 fragments, 199  
 nested elements, 186  
 NEXI, 182  
 overview, 179-180  
 schema, 182  
 schema diversity/heterogeneity, 187  
 structured document retrieval  
 principle, 184  
 tag, 180  
 text-centric, 197
- XML retrieval  
 challenges in, 188  
 context resemblance, 190  
 data-centric, 179, 197  
 evaluation of, 196  
 focused, 199  
 language models *vs.*, 230  
 lexicalized subtrees, 189  
 natural language processing, 230  
 SimNoMerge, computing, 190, 191  
 structural terms, 189  
 text-centric, 197  
 topics in, 193  
 vector space model, 192
- XPath, 181
- Zipf's law, 82-83, 92
- Zone indexes, 107
- Zones, 102-103
- Zone search, 180



025.04 MAN/I

CASMTVK

Books





**I**ntroduction to Information Retrieval is the first textbook with a coherent treatment of classical and web information retrieval, including web search and the related areas of text classification and text clustering. Written from a computer science perspective, it gives an up-to-date treatment of all aspects of the design and implementation of systems for gathering, indexing, and searching documents and of methods for evaluating systems, along with an introduction to the use of machine learning methods on text collections.

Designed as the primary text for a graduate or advanced undergraduate course in information retrieval, the book will also interest researchers and professionals. A complete set of lecture slides and exercises that accompany the book are available on the web.

**Christopher D. Manning** is Associate Professor of Computer Science and Linguistics at Stanford University.

**Prabhakar Raghavan** is Head of Yahoo! Research and a Consulting Professor of Computer Science at Stanford University.

**Hinrich Schütze** is Chair of Theoretical Computational Linguistics at the Institute for Natural Language Processing, University of Stuttgart.

*Cover design by David Levy*

This edition is for sale in South Asia only, not for export elsewhere.

**CAMBRIDGE**  
UNIVERSITY PRESS  
[www.cambridge.org](http://www.cambridge.org)

