Information Retrieval

Implementing and Evaluating Search Engines

Stefan Büttcher
Google Inc.

Charles L. A. Clarke
University of Waterloo

Gordon V. Cormack
University of Waterloo

The MIT Press
Cambridge, Massachusetts
London, England

Typeset by the authors using LaTeX.

Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data
Büttcher, Stefan.
Information retrieval : implementing and evaluating search engines / Stefan Büttcher, Charles L.A. Clarke, and Gordon V. Cormack.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-262-02651-2 (hardcover : alk. paper)
1. Search engines—Programming. 2. Information retrieval. I. Clarke, Charles L. A., 1964–. II. Cormack, Gordon V. III. Title.
TK5105.884.B98 2010
025.5'24—dc22 2009048455

10 9 8 7 6 5 4 3 2 1

Index

α-nDCG, 459
γ code, **193**, 209, 214
δ code, **193–194**, 209
Δ-value, 191
ω code, 194–195

abandonment, 540
abstract data type, 33, 48, 160, 577
accumulator, **145**, 493
accumulator pruning, **147–151**, 493
accuracy, **322**, 335
active learning, 337
Active Server Pages, 510
adaptive algorithm, 38, 62, 65, 304
adaptive coding, 177
adaptive compression, 190
AdaRank-MAP, 399
adhoc task, 24
ADT, *see* abstract data type
advanced search, 160
adversarial IR, 507
alternative hypothesis, 427
anchor text, 277, 507, 513, **536–537**, 555
Antony and Cleopatra, 14
AOL query log, 480
Apache, 27
Arabic, 98
arithmetic coding, **186–189**, 189, 223
arithmetic mean, 44, 68, 409
ASCII, 13, 91
ASP, 510
assessment, 8, 67, 73, **411**, 441–446
average precision, 71, **408**
average response time, 75

B-tree, 113
background language model, 290
backlinks, 534
bag of words, 60, 145, 151, 158
bagging, 376, **385–387**
Basque, 94
batch updates, 229–231
Bayes' Rule, *see* Bayes' Theorem
Bayes' Theorem, 260, 333
Bengali, 97
Bernoulli trials, 418
bias, 413
bigram, 19, 96, 115
binary document format, 11
Binary Independence Model, 261–263
binary interval, 187
binary search, 39, 107, 111, 133, 217, 220
binomial distribution, **418**, 491
bit buffering, 208
blind feedback, *see* pseudo-relevance feedback
blocking, **186**, 216
Bloom filter, 131
BM1, 273
BM11, 273
BM15, 273
BM25, 73, 138–139, 258, **272**, 296, 301, 312, 354
BM25F, 258, **277–279**, 539
body, 193
Bonferroni correction, 428
Bookstein, Abraham, 267
Boolean retrieval, 52, **63–66**, 137, 573
boosting, 376, **387–388**
bootstrap, 424
bootstrap aggregation, 386
Bosak, Jon, 11, 29
Bose-Einstein statistics, 300
bpref, 461
branch prediction, 199, 208, **595**
Bray, Tim, 169
browser extensions, 526
Buckley, Chris, 71
Burrows-Wheeler compression, 191
burst trie, 133

byte-aligned coding, 205–206
bzip2, 191

cache, 41, **479–484**, 544
cache hierarchy, 481
cache line, 594
cache policy, 482–483
candidate phrase, 39
candidate solution, 63
Carnegie Mellon University, 27
case normalization, 84
categorization, 4, **310**, 320, 376
central limit theorem, **431**, 439
CGI, 510
chaining, **107**, 121, 122
Chinese, 95–96, 98
CJK languages, **95–96**
classification, 312, **331–366**
 binary, 331
 decision trees, 360, 364–366
 feature engineering, 338–339
 kernel methods, 357
 linear classifier, 349–353
 multicategory, 388–394
 perceptron algorithm, 352
 probabilistic, 339–349
 Rocchio's method, 354
 SVM, 353
CLEF, 98
Cleverdon, Cyril, 460
clickthrough curve, 540
clickthrough inversion, 540
ClueWeb09 collection, 25
clustering, **4**, 215, 224, 337
code, 177
code tree, **179**, 181, 183
codepoint (Unicode), **91**, 95, 97
combiner, 502
Common Gateway Interface, 510
compression model, 19, **177–180**, 188–190, 196, 202, 361–363
confidence interval, **416–426**, 432, 470
confidence level, 416
contingency table, 332
continuation flag, 205
cosine similarity, **56**, 72
CPU cache, 109, 124, **594**
Cranfield paradigm, 460
Cranfield tests, 460
crawler trap, 557
cross-entropy, **361**, 373
cross-validation, **384–385**, 399
Cutting, Doug, 27, 504
Cutts, Matt, 553

damping factor (PageRank), 518
DDS, *see* deadline-driven scheduling
De Morgan's laws, 65
deadline-driven scheduling, 478
decision tree, 360, **364–366**
DECO algorithm, 539
decoder, 175
decoding performance, **204–209**, 223
decompounding, 94, 98
deep Web, 511
degrees of freedom, 423
delta code, *see* δ code
density function, *see* probability density function
desktop search, 3, 251
DFR, *see* divergence from randomness
dictionary, 33, **106–110**
 hash-based, 107
 sort-based, 107
dictionary compression, 216–222
dictionary group, 217
dictionary interleaving, **114–118**, 221–222
dictionary operations, 106
dictionary-as-a-string approach, 108
diff, 252
digital library, 4
Dijkstra, Edsger, 86
dimensionality reduction, 338, 356
Dirichlet smoothing, 291, 295
disk seek, 111, 145, 233, 238, 493
distribution, 417
 binomial, **418**, 491
 empirical, 419
 exponential, 473
 Gaussian, 417
 geometric, 192, 196, 210
 normal, 417
 Poisson, **267–268**, 473, 548
 Student's t-distribution, 423
 Zipfian, *see* Zipf's law
divergence from randomness, 287, **298–302**
diversity, **455–460**, 537

DMC, 190, 361
docid, 48
docid index, 49
docLeft, 64
docRight, 64
document, **7**–**8**, 45
 element, 7, 564
 update, 7
document format
 binary, 11
 HTML, 9
 Microsoft Office, 13
 OOXML, 13
 PDF, 11
 PostScript, 11
 raw text, 13
 SGML, 11
 XML, 11, 565
document length normalization, 78, 139, **271**–**273**, 295, 296, 299, 301
document map, 105, 214
document partitioning, **490**–**493**, 496
document reordering, 214–216
document structure
 logical, 11
 physical, 11
document type declaration, 570
document type definition, *see* DTD
dot product, 55
double index, 86
DTD, 568–570
duplicate Web page, 549
Dutch, 94, 95, 98
dwell time, 541
dynamic Markov compression, *see* DMC
dynamic page, 510
dynamic programming, 492
dynamic rank, 517, 535

EBCDIC, 91
effect size, 426
effectiveness, 8, 67, 538, 584
efficiency, 8, 75, **468**
eigenvalue, 529
eigenvector, 529
eliteness, **267**, 299, 301
empirical distribution, 419
encoder, 175
English, 95
enterprise search, 4, 511
entropy, 180, 223, **360**
 of English text, 190, 191
 relative, 296
ergodic, 529
Euclidean distance, 56
exhaustivity, 8, **584**
exponential search, *see* galloping search
eXtensible Markup Language, *see* XML

$f_{t,d}$, 48
F-measure, **68**, 371
false negative, 332
false positive, 332
fault tolerance, 472, **496**–**498**
FCFS, *see* first-come first-served
feature engineering, 338–339
feedback, *see* pseudo-relevance feedback
file system search, 3
filtering, 4, 310, 313, 320
 spam, 325, 342
finite-context model, 178, 190
Finnish, 94, 95, 97
FIRE, 98
first, 33
first-come first-served, 473, 474, 478
first-order language model, 19
firstDoc, 49
Fisher, Ronald Aylmer, 414, 427
fixed-effect model, 440
fixed-point iteration, 519
Flash, 13
flat index, *see* schema-independent index
FLWOR expression, 574
follow matrix, 523
forward index, 131
fragmentation, 233, 242
 internal, 108, 123, 124, 483
French, 94, 95
frequency index, 49
front coding, 219–221
function word, 89
fusion, 376, **377**–**381**

galloping search, **42**–**44**, 62, 65, 111, 246
gamma code, *see* γ code
garbage collection, 245–250

Gaussian distribution, 417
GC-list, 160–162
generalizability, 415
generalized concordance list, *see* GC-list
generative model, 286, 289
geometric mean, 44, 409
geometric mean average precision, *see* GMAP
geometric partitioning, 240–242
German, 95, 98
GMAP, 410, 422
Goldilocks, 75
Golomb code, **196–200**, 209
Gosset, William Sealy, 423
GOV2 collection, 25
GPU, 504
graded relevance, *see* relevance
gradient descent, 348
granularity, 112
grouping, 124
gzip, 191

Hadoop, 504
Hamlet, 87, 90
Hamlet, 290, 508
harmonic mean, 68
  weighted, 68
hash table, 107
Hathaway, Anne, 51
heap, 128, **141**, 184
hidden Markov model, 306
hidden Web, 511
HITS algorithm, **532–534**, 554
holdout validation, 383
holistic twig joins, 585
home page finding, 539
host crowding, 493
HTML, 9, 525, 567
HTML anchor, 277, 536
HTML body, 277
HTML header, 277
Huffman code, **181–185**, 189, 200
  canonical, **184–185**, 199, 201
  length-limited, **185**, 201, 209
Hungarian, 94
hybrid index maintenance, 238–239
hyperlinks, 9
HyperText Markup Language, *see* HTML
hypothesis test, 427–429

IDF, *see* inverse document frequency
IE, *see* information extraction
Immediate Merge, **233–235**, 239
impact ordering, **153**, 494
implicit user feedback, 526, 535, **540**, 555
in-degree, 509
incremental crawling, 547
independence assumption, 261
index block size, 116
index construction, 118–131
  in-memory, 119–125
  merge-based, **127–131**, 229
  sort-based, 125–127
  two-pass, 123
index partition, **127**, 228, 240, 471, 488
index pruning, **153–160**, 495
index types, 46–51
index updates
  distributed, 490
  incremental, 231–242
  non-incremental, 243–251
indexable Web, 511
indexing time, 105
Indri, 27–28
INEX, 565, 579
  CAS task, 579
  CO task, 579
infAP, 449–450
inference network model, 280
inferred average precision, *see* infAP
information extraction, 5
information gain, 366
information need, 5
informational query, 514
inner product, 55
Inplace, 236–237
insert-at-back heuristic, 121
inter-query parallelism, 488
interactive search and judging, 443
interpolative coding, **202–204**, 213, 223
intra-query parallelism, 489, 494
intranet, 511
invalidation list, 243–244
inverse document frequency, 57, 264, 581
inverted index, 33
  docid, 49
  frequency, 49
  positional, 49

schema-dependent, 48
schema-independent, 33, 48, 49
Irish, 94
Italian, 94, 95

Japanese, 95, 98
JavaScript, 13
Jelinek-Mercer smoothing, 291, 295
jump vector (PageRank), 523

Kendall's notation, 474
Kendall's τ, 445
Kendall, David, 474
Kendall, Maurice, 445
kernel trick, 358
KL divergence, *see* Kullback-Leibler divergence
Korean, 95
Kullback-Leibler divergence, 156, 286, 296, 527

l_{avg}, 48
l_d, 48
l_t, 34
l_c, 34
LAM, 328
landmark-diff, 252
language model, 17–23
language modeling, 258, 286, **287–298**
Laplace's law of succession, 301
Laplace, Pierre-Simon, 298
last, 33
lastDoc, 49
latency, 8, **470**
Latent Semantic Analysis, 78
lazy evaluation, 244
learning
incremental, 337
on-line, 337
semi-supervised, 336
supervised, 336
transductive, 336
unsupervised, 337
learning to rank, 312, 376, **394–400**
legal search, 46
lemma, 87
lemmatization, 87
length normalization, *see* document length normalization
LETOR, 399
lexeme, 87
LFU, 482
lightweight structure, **160–168**
likelihood ratio, **333**, 341
linear classifier, 349
link analysis, **517–534**, 554
link function, 356
linked list, 122
unrolled, 123, 124, 130
list compression
batched, 196
global, 195
local, 195, 210
ListNet, 399
Little's law, **475**, 476
Little, John, 475
LLRUN, **200–201**, 209, 212, 253
LLRUN-k, 202
log, 422
log-odds, 260
Logarithmic Merge, **240–242**, 249
logical document structure, 11
logistic regression, **346**, 383, 389
gradient descent, 348
multicategory, 392
logit, 260, 422
logit average, 328
long tail, 480, 513
lookup table, 208
Lovins stemmer, 97
LRU, 482
LSA, 78
Lucene, 27

m-cover, 303
M/M/1 queueing model, 475–477
Macbeth, 9, 33, 290, 508, 567, 577
machine learning, 312, 336
macro-average, 322
MAP, **71–74**, 137, 409, 444, 447, 584
MapReduce, **498–503**
Markov chain, **23**, 529
aperiodic, 529
continuous, 475
irreducible, 529
periodic, 529
Markov model, **21–23**, 362
maximal marginal relevance, 461, 493

maximum likelihood, **17**, 289, 297
MaxScore, **143–145**, 491
mean
 arithmetic, 44, 68, 409
 geometric, 44, 409
 harmonic, 68
 weighted harmonic, 68
mean average precision, *see* MAP
mean reciprocal rank, *see* MRR
merge operation
 cascaded, 126, 129
 multiway, 126, 128, 241
meta-analysis, 415, **439–441**
metalanguage, 11
metasearch, 380
micro-average, 322
Microsoft Office, 13
Monty Python's Flying Circus, 78
morphology, 86
move-to-front heuristic, 121
move-to-front pooling, 444
MRR, 322, 409, 539
multicategory classification, 388–394
multicategory ranking, 388–394

N, 48
N_t, 48
n-gram, **92–93**, 95, 96
naïve Bayes, 334
named page finding, 538
navigational query, 513, 539
nDCG, **451–453**, 538
near-duplicate Web page, 549
New Oxford English Dictionary, 160, 169
NEXI, 564, **572–573**
next, 33
nextDoc, 49
NIST, 23
No Merge, **232**, 233
nonparametric code, **192–195**, 216
normal distribution, 417
Normalized Discounted Cumulative Gain, *see* nDCG
novelty, **455–460**, 537, 549
NTCIR, 98
nugget, 459
null hypothesis, 427

Obama, Barack, 441, 515
OCR, *see* optical character recognition, 97
odds, 333
odds ratio, 333
ODP, *see* Open Directory Project
offset, 48
Okapi BM25, *see* BM25
Okapi BM25F, *see* BM25F
omega code, *see* ω code
on-line indexing, *see* index updates
Open Directory Project, 526, 547
open source, 27
Open Source IR Systems, 27–28
optical character recognition, 4, 85
order-preserving, 260
orthography, 94
out-degree, 509
overfitting, **338**, 349
overlap, 580

p-value, 426
Package-Merge, 185
PageRank, 105, **517–532**, 554
 focused, 526
 personalized, 526
 topic-oriented, 526
parametric code, **195–201**, 216
passage retrieval, 302–305
Pat, 160, 169
path expressions, 571
PCA, 554
PDF, 11
Pearson, Karl, 427
per-term index, 112, 133
perceptron algorithm, **352–353**, 357
Perron-Frobenius theorem, 530
phrase search, **35–39**, 111
physical document structure, 11
Pike, Rob, 97
Pinyin, 96
pivoted document length normalization, 78
Poisson distribution, **267–268**, 473, 548
Poisson, Siméon Denis, 268
Polish, 94
pooling (TREC), **73–75**, 411, 441, 443–448
Popper, Karl Raimund, 427
population, 414
Porter stemmer, 87

Porter, Martin, 87, 95
Portuguese, 94
position tree, *see* suffix tree
positional index, 49
postings list, 33, **110–114**, 161
PostScript, 11
power, 406, **434–438**
power method, 530
PPM, 190
pre-allocation
  proportional, 123, 236
pre-allocation factor, 123
preamble, 184, 186, 212, 223
precision, 67–68, 318, 328, **407**
  at k documents, 69, 408
  interpolated, 70
precision of measurement, 413
prefix query, 106, 110, 113, 133
prefix-free, 178
prev, 33
prevDoc, 49
PRF, *see* pseudo-relevance feedback
principal component analysis, 554
prior odds, 334
probabilistic model, 258
probability density function, 341, 417, 473
  cumulative, 417
probability distribution, *see* distribution
Probability Ranking Principle, 8, **259**, 287
proper binary tree, 179
prosecutor's fallacy, 332
proximity ranking, *see* term proximity
PRP, *see* Probability Ranking Principle
pseudo-frequencies, 279
pseudo-relevance feedback, 131, 156, **275–277**, 469

q_t, 51, 271
qrels, 24, **411**, 441, 443
query, 6
query abandonment, 540
query arrival rate, 473
query drift, 277
query execution plan, 244
query expansion, 273, 297
query log, 98, 472, 480, 513
query processing
  document-at-a-time, 139–145
  term-at-a-time, **145–151**, 493
  top-k, 142–145
query reformulation, 540
query term frequency, 271
query time, 105
question answering, 5, 302, 457
queue discipline, 474, 478
queueing theory, 472–477

random access, 35, 111, 116, 196, 216
random error, 413
range encoding, 223
rank effectiveness, 454
rank-biased precision, 461
rank-equivalent, 260
rank-preserving, 260
RankBoost, 399
RankEff, 454
RankSVM, 399
realloc, 123, 124
REBUILD, **229**, 243
recall, 67–68, 88, 138, 318, 328, **407**
recall-precision curve, 70
receptionist, 490
reciprocal rank, 409, 461
reciprocal rank fusion, 380
redundancy, 496–498
refresh policy, 548
region algebra, **160–168**, 169, 567
relevance, **8**, 24, 67, 261, 442
  binary, 8, 407
  graded, 8, 395, **451–453**
relevance feedback, **273–275**, 319, 326, 354
relevance ranking, 3
REMERGE, **229**, 233
replacement algorithm, 482
replication, 471, 497
  dormant, 498
  partial, 497
resampling, 424
research hypothesis, 427
response time, 8, 75, **470**, 476
restart probability (PageRank), 518
retrieval model
  Boolean, 63
  language modeling, 258, 286
  probabilistic, 258
  vector space, 54

retrieval status value, *see* score
Rice code, *see* Golomb code
Robertson, Stephen, 258
Robertson/Spärck Jones weighting formula, 265
robots exclusion protocol, 544
robots.txt, 544
ROC curve, **329**, 397
Rocchio classifier, 354
Rocchio feedback, 280
Romeo and Juliet, 50, 58
routing, 4, 310
RSV, *see* score
run (TREC), 73, 411
Russian, 94

S stemmer, 97
SALSA algorithm, **532–534**, 554
Salton, Gerald, 54
sample, 420
scalar product, 55
scheduling algorithm, 478
schema-dependent index, 48
schema-independent index, 33, 48, 49
score, 7, 59
Scottish Gaelic, 94
seek latency, 493, 592
segmentation, 94, 95
selective dissemination of information, 4, 310
selector, 193, 208
self-indexing, 112
self-information, 299
semi-static coding, 177
semi-supervised learning, 336
sensitivity, 332
SEO, 553
sequential scan, 40
SERP, 510
service rate, *see* throughput
service time, 470
SGML, 11, 568
Shakespeare in Love, 303
Shakespeare, William, 9, 33, 51, 89, 160, 263, 278, 302, 536
Shannon's theorem, *see* source coding th'm
Shannon, Claude, 180, 191
shape property, 141
shard, 500
shingle, 550
shortest job first, 478
signature file, 77, 131
significance, 425
significance level, 416
significance test, 430–438
significant inversion rate, 445
Simple-9, 207
SJF, *see* shortest job first
SMART, 78
smoothing, **20–21**, 264, 290, 340, 450
  Dirichlet, 291, 295
  Jelinek-Mercer, 291, 295
  linear, 291
snippet, 7, 131, 540
  generation, 302
Snowball stemmer, 95, 97
source coding theorem, **180**, 188
source population, 415
Spärck Jones, Karen, 258
spam, 507, 555
spam filtering, 325, 342
Spanish, 94, 95, 98
SPEC, 469
specificity, 8, 332, **584**
spelling correction, 98
splits, 384
stacking, 376, **381–385**
standard error, 386, 422, 429
Standard Generalized Markup Language, *see* SGML
starvation, 479
static coding, 177
static page, 510
static rank, 54, **517–535**
stationary distribution of a Markov chain, 530
steady state, 231, 248, 249
stemming, 84, **86–89**, 95, 97
STL, 120
stochastic matrix, 22
stochastically independent, 268
stopping, 84
stopword, 85, **89–90**
structural index, 585
structural metadata, 568
Student's t-distribution, 423
suffix array, **77**, 131
suffix tree, **77**, 131, 133
summarization, 5

supervised learning, 319, **336**
support vector machine, *see* SVM
support vectors, 353
SVM, **353**, 368
 multicategory, 393
 ranking, 396
SVM^{light}, 397–399
Swedish, 94, 95
symbol, 14, 176
synchronization point, **112**, 195, 219
synonymy, 78
systematic error, 413

t-distribution, 423
table-driven decoding, 208–209
target population, 415
TDT, *see* topic detection and tracking
teleport (PageRank), 523
term, 6, 15
term descriptor, 217
term frequency, 48, **53**, 57, 266
term partitioning, **493**–**495**, 496
term proximity, 54, 60–63, 302–304
term selection value, 274
term vector, 51
test collection, **23**–**26**, 411, 453
 ClueWeb09, 25
 construction, 73–75
 GOV2, 25
 INEX, 583
 TREC45, 26
Text REtrieval Conference, *see* TREC
TF, *see* term frequency
TF-IDF, **57**, 270, 293
The Merry Wives of Windsor, 20
The Winter's Tale, 14
Thompson, Ken, 97
throughput, 8, 75, **470**, 477
token, 13
toolbars, 526
topic, 5
topic detection and tracking, 5
traffic intensity, *see* utilization
transactional query, 514
transductive learning, 336
transfer function, 356, 422
transferability, 415
transition matrix, 22
TREC, **23**–**26**, 67, 98, 272, 410–412
 Filtering Track, 314
 Million Query Track, 25
 Public Spam Corpus, 325
 Robust Track, 282
 Spam Track, 371
 Terabyte Track, 25, 213, 539
TREC45 collection, 26
trigram, 115
true negative, 332
true negative rate, 332
true positive, 332
true positive rate, 332
trust bias, 540
TrustRank, 555
TSV, *see* term selection value
Turkish, 95
twig, 585
two-Poisson model, 267
type 1 error, 331
type 2 error, 331

unary code, 192
Unicode, 13, **91**, 95, 97
universal codeword set, 223
University of Massachusetts, 27
unsupervised learning, 337
URL, 525
user intent, 513
user satisfaction, **410**, 453, 470
UTF-8, 13, **91**, 97
utilization, **476**, 477, 493

validity, 406, 413, **434**–**438**
 external, 415
 internal, 415
vByte, **205**–**206**, 213, 220, 223, 253
vector space model, **54**–**60**, 78
vocabulary, 14

W3C, 564, 572
warmup period (cache), 480
Web
 hidden, 511
 indexable, 511
Web crawler, 507, **541**–**552**, 556
 crawler trap, 557
 incremental, 547

Web graph, 508
Web query, 507, 513
Web search evaluation, 538
Web spam, 507, 555
Webster, John, 18
Wikipedia, 3, 27, 277
Wilcoxon T distribution, 433
Winnie the Pooh, 80
wisdom of crowds, 376
word-aligned coding, 206–207
World Wide Web Consortium, 564
Wumpus, 28

XCG, 584
XML, 11, 29, 160, **565–570**
 declaration, 565
 DTD, 568
 empty-element tag, 566
 exhaustivity, 584
 overlap, 580
 ranked retrieval, 579–584
 specificity, 584
 well-formed document, 568
 XCG, 584
 XML Schema, 568
XML Query, 574
XML Schema, 568, 570
XPath, 564, **571–572**
XQuery, 564, **574–576**
XSL, 572

Zelazny, Roger, 39
zero-order model, 178, 286
Zipf's law, 16, 107, 121, 237, 239, 480, 513
Zipf, George, 16
Ziv-Lempel compression, 191, 220