

Search Engines Information Retrieval in Practice

W. BRUCE CROFT

University of Massachusetts, Amherst

DONALD METZLER

Yahoo! Research

TREVOR STROHMAN

Google Inc.

PEARSON

Boston Columbus Indianapolis New York San Francisco Upper Saddle River
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto
Delhi Mexico City Sao Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Contents

1	Search Engines and Information Retrieval	1
1.1	What Is Information Retrieval?	1
1.2	The Big Issues	4
1.3	Search Engines	6
1.4	Search Engineers	9
2	Architecture of a Search Engine	13
2.1	What Is an Architecture?	13
2.2	Basic Building Blocks	14
2.3	Breaking It Down	17
2.3.1	Text Acquisition	17
2.3.2	Text Transformation	19
2.3.3	Index Creation	22
2.3.4	User Interaction	23
2.3.5	Ranking	25
2.3.6	Evaluation	27
2.4	How Does It <i>Really</i> Work?	28
3	Crawls and Feeds	31
3.1	Deciding What to Search	31
3.2	Crawling the Web	32
3.2.1	Retrieving Web Pages	33
3.2.2	The Web Crawler	35
3.2.3	Freshness	37
3.2.4	Focused Crawling	41
3.2.5	Deep Web	41

3.2.6	Sitemaps	43
3.2.7	Distributed Crawling	44
3.3	Crawling Documents and Email	46
3.4	Document Feeds	47
3.5	The Conversion Problem	49
3.5.1	Character Encodings	50
3.6	Storing the Documents	52
3.6.1	Using a Database System	53
3.6.2	Random Access	53
3.6.3	Compression and Large Files	54
3.6.4	Update	56
3.6.5	BigTable	57
3.7	Detecting Duplicates	60
3.8	Removing Noise	63
4	Processing Text	75
4.1	From Words to Terms	75
4.2	Text Statistics	77
4.2.1	Vocabulary Growth	82
4.2.2	Estimating Collection and Result Set Sizes	85
4.3	Document Parsing	88
4.3.1	Overview	88
4.3.2	Tokenizing	89
4.3.3	Stopping	92
4.3.4	Stemming	93
4.3.5	Phrases and N-grams	99
4.4	Document Structure and Markup	103
4.5	Link Analysis	106
4.5.1	Anchor Text	107
4.5.2	PageRank	107
4.5.3	Link Quality	113
4.6	Information Extraction	115
4.6.1	Hidden Markov Models for Extraction	117
4.7	Internationalization	120

5	Ranking with Indexes	127
5.1	Overview	127
5.2	Abstract Model of Ranking	128
5.3	Inverted Indexes	131
5.3.1	Documents	133
5.3.2	Counts	135
5.3.3	Positions	136
5.3.4	Fields and Extents	138
5.3.5	Scores	140
5.3.6	Ordering	141
5.4	Compression	142
5.4.1	Entropy and Ambiguity	144
5.4.2	Delta Encoding	146
5.4.3	Bit-Aligned Codes	147
5.4.4	Byte-Aligned Codes	150
5.4.5	Compression in Practice	151
5.4.6	Looking Ahead	153
5.4.7	Skipping and Skip Pointers	153
5.5	Auxiliary Structures	156
5.6	Index Construction	158
5.6.1	Simple Construction	158
5.6.2	Merging	159
5.6.3	Parallelism and Distribution	160
5.6.4	Update	166
5.7	Query Processing	167
5.7.1	Document-at-a-time Evaluation	168
5.7.2	Term-at-a-time Evaluation	170
5.7.3	Optimization Techniques	172
5.7.4	Structured Queries	180
5.7.5	Distributed Evaluation	182
5.7.6	Caching	183
6	Queries and Interfaces	191
6.1	Information Needs and Queries	191
6.2	Query Transformation and Refinement	194
6.2.1	Stopping and Stemming Revisited	194
6.2.2	Spell Checking and Suggestions	197

6.2.3	Query Expansion	203
6.2.4	Relevance Feedback	212
6.2.5	Context and Personalization	215
6.3	Showing the Results	219
6.3.1	Result Pages and Snippets	219
6.3.2	Advertising and Search	222
6.3.3	Clustering the Results	225
6.4	Cross-Language Search	230
7	Retrieval Models	237
7.1	Overview of Retrieval Models	237
7.1.1	Boolean Retrieval	239
7.1.2	The Vector Space Model	241
7.2	Probabilistic Models	247
7.2.1	Information Retrieval as Classification	248
7.2.2	The BM25 Ranking Algorithm	254
7.3	Ranking Based on Language Models	256
7.3.1	Query Likelihood Ranking	258
7.3.2	Relevance Models and Pseudo-Relevance Feedback	265
7.4	Complex Queries and Combining Evidence	271
7.4.1	The Inference Network Model	272
7.4.2	The Galago Query Language	277
7.5	Web Search	283
7.6	Machine Learning and Information Retrieval	287
7.6.1	Learning to Rank	288
7.6.2	Topic Models and Vocabulary Mismatch	292
7.7	Application-Based Models	295
8	Evaluating Search Engines	301
8.1	Why Evaluate?	301
8.2	The Evaluation Corpus	303
8.3	Logging	309
8.4	Effectiveness Metrics	312
8.4.1	Recall and Precision	312
8.4.2	Averaging and Interpolation	317
8.4.3	Focusing on the Top Documents	322
8.4.4	Using Preferences	325

8.5	Efficiency Metrics	326
8.6	Training, Testing, and Statistics	329
8.6.1	Significance Tests	329
8.6.2	Setting Parameter Values	334
8.6.3	Online Testing	336
8.7	The Bottom Line	337
9	Classification and Clustering	343
9.1	Classification and Categorization	344
9.1.1	Naïve Bayes	346
9.1.2	Support Vector Machines	355
9.1.3	Evaluation	363
9.1.4	Classifier and Feature Selection	363
9.1.5	Spam, Sentiment, and Online Advertising	368
9.2	Clustering	377
9.2.1	Hierarchical and K -Means Clustering	379
9.2.2	K Nearest Neighbor Clustering	388
9.2.3	Evaluation	390
9.2.4	How to Choose K	391
9.2.5	Clustering and Search	393
10	Social Search	401
10.1	What Is Social Search?	401
10.2	User Tags and Manual Indexing	404
10.2.1	Searching Tags	406
10.2.2	Inferring Missing Tags	408
10.2.3	Browsing and Tag Clouds	410
10.3	Searching with Communities	412
10.3.1	What Is a Community?	412
10.3.2	Finding Communities	413
10.3.3	Community-Based Question Answering	419
10.3.4	Collaborative Searching	424
10.4	Filtering and Recommending	427
10.4.1	Document Filtering	427
10.4.2	Collaborative Filtering	436
10.5	Peer-to-Peer and Metasearch	442
10.5.1	Distributed Search	442

10.5.2 P2P Networks	446
11 Beyond Bag of Words	455
11.1 Overview	455
11.2 Feature-Based Retrieval Models	456
11.3 Term Dependence Models	458
11.4 Structure Revisited	463
11.4.1 XML Retrieval	465
11.4.2 Entity Search	468
11.5 Longer Questions, Better Answers	470
11.6 Words, Pictures, and Music	474
11.7 One Search Fits All?	483
References	491
Index	517