

Contents

<i>Figures</i>	<i>page xi</i>
<i>Foreword by C. J. van Rijsbergen</i>	<i>xv</i>
<i>Preface</i>	<i>xix</i>

1 Overview	1
1.1 Finding Out About – A Cognitive Activity	1
1.1.1 Working within the IR Tradition	8
1.2 Keywords	10
1.2.1 Elements of the Query Language	10
1.2.2 Topical Scope	11
1.2.3 Document Descriptors	12
1.3 Query Syntax	13
1.3.1 Query Sessions	14
1.4 Documents	16
1.4.1 Structured Aspects of Documents	19
1.4.2 Corpora	20
1.4.3 Document Proxies	20
1.4.4 Genre	21
1.4.5 Beyond Text	22
1.5 Indexing	26
1.5.1 Automatically Selecting Keywords	27
1.5.2 Computer-Assisted Indexing	28

1.6 FOA versus Database Retrieval	29
1.7 How Well Are We Doing?	34
1.8 Summary	36
2 Extracting Lexical Features	39
2.1 Building Useful Tools	39
2.2 Interdocument Parsing	40
2.3 Intradocument Parsing	42
2.3.1 Stemming and Other Morphological Processing	44
2.3.2 Noise Words	47
2.3.3 Summary	48
2.4 Example Corpora	48
2.5 Implementation	50
2.5.1 Basic Algorithm	51
2.5.2 Fine Points	54
2.5.3 Software Libraries	58
3 Weighting and Matching against Indices	60
3.1 Microscopic Semantics and the Statistics of Communication	60
3.2 Remember Zipf	62
3.2.1 Looking for Meaning in All the Wrong Places (At the Character Level)	63
3.2.2 Zipf's Own Explanation	64
3.2.3 Benoit Mandelbrot's Explanation	66
3.2.4 Herbert Simon's Explanation	67
3.2.5 More Recent Zipfian Sightings	68
3.2.6 Summary	71
3.3 A Statistical Basis for Keyword Meaning	71
3.3.1 Lexical Consequences, Internal/External Perspectives	71
3.3.2 Word Occurrence as a Poisson Process	73
3.3.3 Resolving Power	76
3.3.4 Language Distribution	78
3.3.5 Weighting the <i>Index</i> Relation	81
3.3.6 Informative Signals versus Noise Words	83
3.3.7 Inverse Document Frequency	84
3.4 Vector Space	86

3.4.1	Keyword Discrimination	88
3.4.2	Vector Length Normalization	89
3.4.3	Summary: SMART Weighting Specification	92
3.5	Matching Queries against Documents	93
3.5.1	Measures of Association	94
3.5.2	Cosine Similarity	95
3.6	Calculating TF-IDF Weighting	96
3.7	Computing Partial Match Scores	97
3.8	Summary	101
4	Assessing the Retrieval	105
4.1	Personal Assessment of Relevance	106
4.1.1	Cognitive Assumptions	106
4.2	Extending the Dialog with <i>RelFbk</i>	109
4.2.1	Using <i>RelFbk</i> for Query Refinement	111
4.2.2	Using <i>RelFbk</i> to Adapt Documents' Indices	115
4.2.3	Summary	116
4.3	Aggregated Assessment: Search Engine Performance	116
4.3.1	Underlying Assumptions	116
4.3.2	Consensual Relevance	118
4.3.3	Traditional Evaluation Methodologies	119
4.3.4	Basic Measures	122
4.3.5	Ordering the <i>Retr</i> Set	124
4.3.6	Normalized Recall and Precision	128
4.3.7	Multiple Retrievals across Varying Queries	129
4.3.8	One-Parameter Criteria	132
4.3.9	Test Corpora	135
4.3.10	Other Measures	137
4.4	RAVE: A Relevance Assessment VEHICLE	141
4.4.1	RAVeUnion	141
4.4.2	RAVePlan	142
4.4.3	Interactive RAVE	143
4.4.4	RAVeCompile	145
4.5	Summary	146
5	Mathematical Foundations	149
5.1	Derivation of Zipf's Law for Random Texts	149
5.1.1	Discussion	152

5.2	Dimensionality Reduction	153
5.2.1	A Simple Example	153
5.2.2	Formal Notions of Similarity	155
5.2.3	Singular Value Decomposition	156
5.2.4	How Many Dimensions k to Reduce to?	157
5.2.5	Other Uses of Vector Space	158
5.2.6	Computational Considerations	159
5.2.7	“Latent Semantic” Claims	159
5.3	Preference Relations	161
5.3.1	Multidimensional Scaling	161
5.3.2	Information in <i>RelFbk</i>	163
5.3.3	Connections between MDS and LSI	165
5.4	Clustering	165
5.4.1	The Cluster Hypothesis	165
5.4.2	Clustering Algorithms	166
5.5	Probabilistic Retrieval	167
5.5.1	Probability Ranking Principle	168
5.5.2	Bayesian Inversion	169
5.5.3	Odds Calculation	170
5.5.4	Binary Independence Model	170
5.5.5	Linear Discriminators	173
5.5.6	Cost Analysis	175
5.5.7	Bayesian Networks	175
6	Inference beyond the <i>Index</i>	182
6.1	Citation: Interdocument Links	185
6.1.1	Bibliometric Analysis of Science	187
6.1.2	Time Scale	190
6.1.3	Legal Citation	191
6.1.4	Citations and Arguments	193
6.1.5	Analyzing WWW Adjacency	195
6.2	Hypertext, Intradocument Links	199
6.2.1	Footnotes, Hyperfootnotes, and cf.	200
6.2.2	Hierarchic Containment	201
6.2.3	Argument Relations	205
6.2.4	Intra- versus Interdocument Relations	207
6.2.5	Beyond Unary <i>About</i> (k) Predicates	209
6.3	Keyword Structures	210

6.3.1 Automatic Thesaurus Construction	211
6.3.2 Corpus-Based Linguistics and WordNet	213
6.3.3 Taxonomies	217
6.4 Social Relations among Authors	220
6.4.1 AI Genealogy	221
6.4.2 An Empirical Foundation for a Philosophy of Science	223
6.5 Modes of Inference	224
6.5.1 Theorem-Proving Models for Relevance	224
6.5.2 Spreading Activation Search	225
6.5.3 Discovering Latent Knowledge within a Corpus	234
6.6 Deep Interfaces	238
6.6.1 Geographical Hitlists	238
6.7 FOA (The Law)	242
6.8 FOA (Evolution)	245
6.9 Text-Based Intelligence	247
7 Adaptive Information Retrieval	252
7.1 Background	252
7.1.1 Training against Manual Indices	254
7.1.2 Alternative Tasks for Learning	255
7.1.3 Sources of Feedback	256
7.2 Building Hypotheses <u>about</u> Documents	258
7.2.1 Feature Selection	259
7.2.2 Hypothesis Spaces	262
7.3 Learning Which Documents to Route	263
7.3.1 Widrow-Hoff	265
7.3.2 User Drift and Event Tracking	266
7.4 Classification	267
7.4.1 Modeling Documents	269
7.4.2 Training a Classifier	271
7.4.3 Priors	272
7.5 Other Approaches to Classification	273
7.5.1 Nearest-Neighbor Matching	273
7.5.2 Boolean Predicates	273
7.5.3 When Irrelevant Attributes Abound	274
7.5.4 Combining Classifiers	275
7.5.5 Hierarchic Classification	277

7.6	Information-Seeking Agents	279
7.6.1	Exploiting Linkage for Context	279
7.6.2	The InfoSpiders Algorithm	281
7.6.3	Adapting to “Spatial” Context	284
7.7	Other Learning Applications and Issues	286
7.7.1	Adaptive Lenses	286
7.7.2	Adapting to Fluid Language Use	288
7.8	Symbolic and Subsymbolic Learning	288
8	Conclusions and Future Directions	292
8.1	Things that Are Changing	292
8.1.1	WWW Crawling	294
8.2	Things that Stay the Same	298
8.2.1	The FOA Language Game	299
8.2.2	Sperber and Wilson’s “Relevance”	304
8.2.3	Argument Structures	305
8.2.4	User as Portal	306
8.3	Who Needs to FOA	307
8.3.1	Authors	308
8.3.2	Scientists	309
8.3.3	The Changing Economics of Publishing	311
8.3.4	Teachers and Students	313
8.4	Summary	316
	<i>(Active) Colophon</i>	318
	<i>Bibliography</i>	321
	<i>Index</i>	347

Figures

Finding Out About	<i>page</i> 1
1.1 The FOA Conversation Loop	5
1.2 Retrieval of Documents in Response to a Query	7
1.3 Assessment of the Retrieval	8
1.4 Schematic of Search Engine	10
1.5 A Query Session	15
1.6 Finding Out About POLITICAL FIDELITY	24
1.7 Obsolete Concert Schedule	25
1.8 Results of SCSI Search of UseNet	31
1.9 A Relevant Posting	31
1.10 Comparison of Retrieved versus Relevant Documents	35
2.1 Parsing Email and AIT to Common Specifications	42
2.2 Finite State Machine	43
2.3 AIT Year Distribution	49
2.4 Basic Postings Data Structures	53
2.5 Refined Postings Data Structures	54
2.6 STAIRS Posting Information	56
2.7 Quoted Lines in an Email Message	57
3.1 Zipfian Distribution of AIT Words	63
3.2 Rank/Frequency Distribution of Click-Paths	70
3.3 Resolving Power	77
3.4 Specificity/Exhaustivity Trade-Offs	78

3.5	Indexing Graph	80
3.6	Hypothetical Word Distributions	84
3.7	Vector Space	87
3.8	Length Normalization of Vector Space	90
3.9	Sensitivity of IDF to “Document” Size	90
3.10	Pivot-Based Document Length Normalization	91
4.1	Relevance Scale	108
4.2	<i>RelFbk</i> Labeling of the <i>Retr</i> Set	109
4.3	Query Session, Linked by <i>RelFbk</i>	110
4.4	<i>RelFbk</i> Labels in Vector Space	112
4.5	Various Ways of Being Irrelevant	112
4.6	Using <i>RelFbk</i> to Refine the Query	113
4.7	Document Modifications due to <i>RelFbk</i>	115
4.8	Consensual Relevance	119
4.9	Relevant versus Retrieved Sets	122
4.10	Recall/Precision Curve	127
4.11	Instability of Beginning of Re/Pre Curve	127
4.12	Best/Worst Retrieval Envelope	128
4.13	Normalized Recall	129
4.14	Multiple Queries, Fixed Recall Levels	130
4.15	11-Point Average Re/Pre Curves	131
4.16	TREC Query	136
4.17	Distinguishing between Overlapping Distributions	139
4.18	Operating Characteristic Curve	140
4.19	RAVE Interface	144
5.1	Lexicographic Tree Underlying Zipfian Distribution	150
5.2	α as Function of M , Number of Distinct Characters	152
5.3	Weight and Height Data Reduction	154
5.4	SVD Decomposition	157
5.5	Retrieval Performance as Function of SVD Dimension (k)	158
5.6	Random Variables Underlying Binary Independence Model	172
5.7	Interaction between Parental Influences	176
5.8	Bayesian Network Representation for FOA	177
5.9	Concept-Matching Version of Bayesian Network	178
5.10	Two Examples of Query Modeling	179
6.1	Other Information Available for FOA	184
6.2	Basic Structure of Citations	186

6.3	Temporal Structure in Citations	188
6.4	Rose's Two-Dimensional Analysis of Shepherd Treatment Codes	194
6.5	Citation-Expanded Hitlist	198
6.6	Containment and Document References	202
6.7	Correlation of Passages	203
6.8	Visualizing Topical Distributions	204
6.9	Topical Document Distributions	205
6.10	Intradocument Relations	206
6.11	FOA Overview	206
6.12	Typography Used to Isolate Talmudic Voices	208
6.13	HTML Version of the Talmud	209
6.14	Example of the Term LYMPHOMA within the MeSH Keyword Thesaurus	212
6.15	Bipolar Organization of Adjectives in WordNet	217
6.16	Nesting Taxonomies from Various Sources	218
6.17	A Sample of the AI Genealogical Record	222
6.18	Spreading Activation Search	226
6.19	Subnet Corresponding to Each Document	227
6.20	AIR Interface	229
6.21	Semantic Net of Legal Document Relations	234
6.22	Swanson's Search for Latent Knowledge	236
6.23	CIVIL WAR BATTLE Query, Standard Textual Hitlist	241
6.24	CIVIL WAR BATTLE Query, Geographical Presentation	242
6.25	The Long Life of Legal Documents	243
6.26	The Annotation Relation between Text and Sequence Data	246
7.1	Learning Conceptual Structures	253
7.2	Browsing across Queries in Same Session	257
7.3	Inductive Bias	263
7.4	Document Modifications due to <i>RelFbk</i>	264
7.5	Training a Classifier	268
7.6	RIPPER Classification Rule	273
7.7	Covering Algorithms	274
7.8	Combining Experts	275
7.9	Optimal Weightings Distribution	276
7.10	Hierarchic Classification	278
7.11	Ancestors of a Class	279

7.12	Adaptive Lens	287
8.1	Crawler Coverage	295
8.2	A Semiotic View of FOA	301
8.3	A Semiotic Analysis of Keyword Mismatch	302
8.4	Grice's Maxims	303
8.5	Query as Portal, Connecting Corpora	307
8.6	The Tell/Ask Duality	314