

FINDING OUT ABOUT:  
A Cognitive Perspective on  
Search Engine Technology and the  
World Wide Web

*Final Draft: 28 January 2000*

©Richard K. Belew



# Contents

<b>1</b>	<b>Overview</b>	<b>1</b>
1.1	Finding Out About - a cognitive activity . . . . .	1
1.1.1	Working within the IR Tradition . . . . .	7
1.2	Keywords . . . . .	8
1.2.1	Elements of the query language . . . . .	9
1.2.2	Topical scope . . . . .	9
1.2.3	Document descriptors . . . . .	10
1.3	Query syntax . . . . .	11
1.3.1	Query sessions . . . . .	12
1.4	Documents . . . . .	13
1.4.1	Structured aspects of documents . . . . .	15
1.4.2	Corpora . . . . .	16
1.4.3	Document proxies . . . . .	16
1.4.4	Genre . . . . .	17
1.4.5	Beyond text . . . . .	18
1.5	Indexing . . . . .	21
1.5.1	Automatically selecting keywords . . . . .	21
1.5.2	Computer-assisted indexing . . . . .	22
1.6	FOA versus database retrieval . . . . .	23
1.7	How well are we doing? . . . . .	27
1.8	Summary . . . . .	29
<b>2</b>	<b>Extracting Lexical Features</b>	<b>31</b>
2.1	Building useful tools . . . . .	31
2.2	Inter-document parsing . . . . .	32
2.3	Intra-document parsing . . . . .	34
2.3.1	Stemming and other morphological processing . . . . .	35
2.3.2	Noise words . . . . .	37
2.3.3	Summary . . . . .	38
2.4	Example corpora . . . . .	39
2.5	Implementation . . . . .	40
2.5.1	Basic algorithm . . . . .	41
2.5.2	Fine points . . . . .	43
2.5.3	Software libraries . . . . .	46

<b>3</b>	<b>Weighting and matching against Indices</b>	<b>49</b>
3.1	Microscopic semantics and the statistics of communication . . . .	49
3.2	Remember Zipf . . . . .	51
3.3	A statistical basis for keyword meaning . . . . .	59
3.3.1	Lexical consequences, internal/external perspectives . . .	59
3.3.2	Word occurrence as a Poisson process . . . . .	62
3.3.3	Resolving power . . . . .	64
3.3.4	Language distribution . . . . .	66
3.3.5	Weighting the index relation . . . . .	69
3.3.6	Informative signals versus noise words . . . . .	70
3.3.7	Inverse document frequency . . . . .	72
3.4	Vector space . . . . .	73
3.4.1	Keyword discrimination . . . . .	74
3.4.2	Vector length normalization . . . . .	75
3.4.3	Summary: SMART weighting specification . . . . .	78
3.5	Matching queries against documents . . . . .	79
3.5.1	Measures of association . . . . .	80
3.5.2	Cosine similarity . . . . .	81
3.6	Calculating TF-IDF Weighting . . . . .	81
3.7	Computing partial match scores . . . . .	82
3.8	Conclusion . . . . .	86
<b>4</b>	<b>Assessing the retrieval</b>	<b>89</b>
4.1	INDIVIDUAL'S assessment of relevance . . . . .	90
4.1.1	Cognitive assumptions . . . . .	90
4.2	Extending the dialog with <i>RelFbk</i> . . . . .	92
4.2.1	Using <i>RelFbk</i> for query refinement . . . . .	94
4.2.2	Using <i>RelFbk</i> to adapt documents' indices . . . . .	97
4.2.3	Summary . . . . .	98
4.3	INDIVIDUALS' assessment: search engine performance . . . . .	98
4.3.1	Underlying assumptions . . . . .	99
4.3.2	Consensual relevance . . . . .	100
4.3.3	Traditional evaluation methodologies . . . . .	101
4.3.4	Basic measures . . . . .	103
4.3.5	Ordering the <i>Retr</i> set . . . . .	105
4.3.6	Normalized recall and precision . . . . .	107
4.3.7	Multiple retrievals across varying queries . . . . .	109
4.3.8	One-parameter criteria . . . . .	112
4.3.9	Test corpora . . . . .	114
4.3.10	Other measures . . . . .	116
4.4	RAVe: A Relevance Assessment VEHICLE . . . . .	119
4.4.1	RAVeUnion . . . . .	119
4.4.2	RAVePlan . . . . .	120
4.4.3	Interactive RAVE . . . . .	121
4.4.4	RAVeCompile . . . . .	121
4.5	Summary . . . . .	123

<b>5</b>	<b>Mathematical Foundations</b>	<b>125</b>
5.1	Derivation of Zipf’s Law for random texts . . . . .	125
5.1.1	Discussion . . . . .	128
5.2	Dimensionality reduction . . . . .	129
5.2.1	A simple example . . . . .	130
5.2.2	Formal notions of similarity . . . . .	131
5.2.3	Singular value decomposition . . . . .	131
5.2.4	How many dimensions $k$ to reduce to? . . . . .	133
5.2.5	Other uses of vector space . . . . .	133
5.2.6	Computational considerations . . . . .	134
5.2.7	“Latent semantic” claims . . . . .	135
5.3	Preference Relations . . . . .	136
5.3.1	Multidimensional scaling . . . . .	136
5.3.2	Information in <i>RelFbk</i> . . . . .	138
5.3.3	Connections between MDS and LSI . . . . .	139
5.4	Clustering . . . . .	140
5.4.1	The Clustering Hypothesis . . . . .	140
5.4.2	Clustering algorithms . . . . .	140
5.5	Probabilistic retrieval . . . . .	142
5.5.1	Probability Ranking Principle . . . . .	142
5.5.2	Bayesian inversion . . . . .	143
5.5.3	Odds calculation . . . . .	143
5.5.4	Binary Independence Model . . . . .	144
5.5.5	Linear discriminators . . . . .	146
5.5.6	Cost analysis . . . . .	147
5.5.7	Bayesian networks . . . . .	147
<b>6</b>	<b>Inference beyond the <i>Index</i></b>	<b>153</b>
6.1	Citation: inter-document links . . . . .	156
6.1.1	Bibliometric analysis of science . . . . .	157
6.1.2	Time-scale . . . . .	160
6.1.3	Legal citation . . . . .	161
6.1.4	Discussion . . . . .	163
6.1.5	Analyzing WWW adjacency . . . . .	165
6.2	Hypertext, intra-document links . . . . .	169
6.2.1	Footnotes, hyper-footnotes and <i>cf.</i> . . . . .	169
6.2.2	Hierarchic containment . . . . .	170
6.2.3	Argument relations . . . . .	175
6.2.4	Intra- vs. inter-document relations . . . . .	175
6.2.5	Beyond unary <i>About(k)</i> predicates . . . . .	179
6.3	Keyword structures . . . . .	179
6.3.1	Automatic thesaurus construction . . . . .	180
6.3.2	Corpus-based linguistics and WordNet . . . . .	180
6.3.3	Taxonomies . . . . .	184
6.4	Social relations among authors . . . . .	187
6.4.1	A.I. Geneology . . . . .	187

6.4.2	An emprical foundation for a philosophy of Science . . . . .	189
6.5	Modes of inference . . . . .	190
6.5.1	Theorem-proving models for relevance . . . . .	190
6.5.2	Spreading activation search . . . . .	191
6.5.3	Discovering latent knowledge within a corpus . . . . .	198
6.6	Deep interfaces . . . . .	202
6.6.1	Geographical hitlists . . . . .	202
6.7	FOA(The Law) . . . . .	206
6.8	FOA(Evolution) . . . . .	208
6.9	Text-based intelligence . . . . .	210
<b>7</b>	<b>Adaptive Information Retrieval</b>	<b>213</b>
7.1	Background . . . . .	213
7.1.1	Training against manual indices . . . . .	215
7.1.2	Alternative Tasks for Learning . . . . .	215
7.1.3	Sources of Feedback . . . . .	216
7.2	Building hypotheses about documents . . . . .	218
7.2.1	Feature selection . . . . .	219
7.2.2	Hypothesis spaces . . . . .	221
7.3	Learning which documents to route . . . . .	223
7.3.1	Widrow-Hoff . . . . .	224
7.3.2	User drift and event tracking . . . . .	225
7.4	Classification . . . . .	225
7.4.1	Modeling documents . . . . .	228
7.4.2	Training a classifier . . . . .	229
7.4.3	Priors . . . . .	229
7.5	Other approaches to classification . . . . .	230
7.5.1	Nearest-neighbor Matching . . . . .	230
7.5.2	Boolean predicates . . . . .	230
7.5.3	When irrelevant attributes abound . . . . .	231
7.5.4	Combining Classifiers . . . . .	232
7.5.5	Hierarchic classification . . . . .	234
7.6	Information-seeking agents . . . . .	235
7.6.1	Exploiting linkage for context . . . . .	235
7.6.2	The InfoSpiders algorithm . . . . .	237
7.6.3	Adapting to “spatial” context . . . . .	239
7.7	Other learning applications and issues . . . . .	240
7.7.1	Adaptive Lenses . . . . .	240
7.7.2	Adapting to fluid language use . . . . .	242
7.8	Symbolic and Subsymbolic Learning . . . . .	243
<b>8</b>	<b>Conclusions &amp; Future Directions</b>	<b>245</b>
8.1	Things that are changing . . . . .	245
8.1.1	WWW crawling . . . . .	247
8.2	Things that stay the same . . . . .	250
8.2.1	The FOA language game . . . . .	250

8.2.2	Sperber & Wilson’s “relevance” . . . . .	254
8.2.3	Argument Structures . . . . .	255
8.2.4	User as portal . . . . .	256
8.3	Who needs to FOA . . . . .	257
8.3.1	Authors . . . . .	257
8.3.2	Scientists . . . . .	258
8.3.3	The changing economics of publishing . . . . .	260
8.3.4	Teachers and students . . . . .	261
8.4	Summary . . . . .	264
	<b>(Active) Colophon</b>	<b>265</b>