

CONTENTS IN DETAIL

INTRODUCTION

xvii

PART I AN INTRODUCTION TO SPAM FILTERING

1	THE HISTORY OF SPAM	3
	The Definition of Spam	4
	The Very First Spam	4
	Spam: The Early Years	7
	Jay-Jay's College Fund	7
	The Jesus Spam	9
	Canter & Siegel	10
	Cancelmoose	13
	Jeff Slaton, the "Spam King"	14
	"Krazy" Kevin Lipsitz	15
	Stanford Wallace, Cyber Promotions	15
	Floodgate—The First Spamware	16
	Other Significant Events in 1995	16
	War Waged on Spam	17
	Spamhaus	17
	Unsolicited Commercial Email	19
	Spam Out of Control	19
	1998, 1999, and 2000: Three Years of War on Spam	20
	Network Solutions	22
	2001 to the Present: Exponential Spam Growth	22
	Final Thoughts	23
2	HISTORICAL APPROACHES TO FIGHTING SPAM	25
	Primitive Language Analysis	26
	Blacklisting	27
	Propagation and Maintenance Problems	28
	Heuristic Filtering	29
	Brightmail	29
	SpamAssassin	30
	Drawbacks to Heuristic Filtering	31
	Maintenance Headaches	31
	Scoring	32

Whitelisting	32
A Little Too Effective	32
Forgeries	33
Challenge/Response	34
Problems with Challenge/Response	34
Throttling	35
TarProxy	35
Other Throttling Tools	36
Collaborative Filtering	37
Address Obfuscation	38
New Standards	39
Authenticated SMTP	39
Sender Policy Framework	40
Litigation	41
Spammer Fingerprinting	43
Intellectual Property	44
Final Thoughts	44

3

LANGUAGE CLASSIFICATION CONCEPTS 45

Understanding Accuracy	46
Machine Learning	46
Concept Learning	47
Using Language Classification to Fight Spam	47
Training	48
Statistical Filtering and Bayesian Analysis	49
Components of a Language Classifier	49
The Historical Dataset	50
The Tokenizer	51
The Analysis Engine	53
Providing Feedback	54
Training	55
Train-Everything (TEFT)	55
Train-on-Error (TOE)	56
Train-Until-Mature (TUM)	56
Train-Until-No-Errors (TUNE)	57
When to Train	57
An Example of a Filter Instance	58
Step 1: Tokenize the Message	58
Step 2: Build a Decision Matrix	59
Step 3: Evaluate the Decision Matrix	59
Step 4: Train the Message	60
Step 5: Correct Errors	60
Efficacy of Statistical Filtering	60
The Future of Language Classification	61
The Sovereignty of Statistical Filtering	61
Final Thoughts	62

4	STATISTICAL FILTERING FUNDAMENTALS	63
An Imperfect Solution		64
Building a Historical Dataset		65
Corpus Feeding		65
Starting from Scratch		66
Correcting Errors		67
The Tokenizer and Calculating Token Values		68
Single-Corpus Tokens		70
A Biased Filter		71
Hapaxes		71
Final Product		72
The Analysis Engine		72
Sorting		73
Statistical Combination		74
Bayesian Combination (Paul Graham)		75
Bayesian Combination (Brian Burton)		76
Robinson's Geometric Mean Test		78
Fisher-Robinson's Inverse Chi-Square		79
Improvements to Statistical Analysis		80
Improving the Decision Matrix		80
Improvements to Tokenization		81
Statistical Sedation		81
Iterative Training		82
Learning New Tricks		83
Final Thoughts		83

PART II FUNDAMENTALS OF STATISTICAL FILTERING

5	DECODING: UNCOMBOBULATING MESSAGES	87
Introduction to Encoding		88
Decoding		88
Message Body Encodings		89
Quoted-Printable Encoding		91
Base64 Encoding		91
Custom Encodings		92
Message Header Encodings		92
HTML Encodings		93
Message Actualization		94
Supporting Software		95
Final Thoughts		95

6 **TOKENIZATION: THE BUILDING BLOCKS OF SPAM** **97**

Tokenizing a Heuristic Function	98
Basic Delimiters	98
Redundancy	99
Other Delimiters	100
Exceptions	101
Token Reassembly	101
Degeneration	102
Header Optimizations	103
URL Optimizations	104
HTML Tokenization	105
Word Pairs	107
Sparse Binary Polynomial Hashing	108
Internationalization	108
Final Thoughts	109

7 **THE LOW-DOWN DIRTY TRICKS OF SPAMMERS** **111**

Successful Filtering	112
No More Headaches	112
A Weak Link in Statistical Filters?	113
Attacks on Tokenizers	113
Encoding Abuses	114
Header Encodings	114
Hypertextus Interruptus	115
ASCII Spam	117
Text-Splitting	119
Table-Based Obfuscation	121
URL Encodings	123
Symbolic Text	124
Just Plain Dumb	124
Attacks on the Dataset	125
Mailing List Attacks	126
Bayesian Poisoning	127
Empty but Not Empty Probes	130
Attacks on the Decision Matrix	132
Image Spams	132
Random Strings of Text	134
Word Salad	135
Directed Attacks	137
Final Thoughts	139

8 **DATA STORAGE FOR A ZILLION RECORDS** **141**

Storage Considerations	142
Disk Space	142
Speed	142

Locking	143
Portability	143
Statefulness	143
Recovery	143
I/O Contention	144
Random-Access Features	144
Ease of Use	144
Storage Framework	145
Third-Party Storage Solutions	147
Stateless Database Implementations	147
Stateful SQL-Based Solutions	149
Peter Graf's PBL ISAM Library	151
SQLite	153
Proprietary Implementations	155
Final Thoughts	155

9 SCALING IN LARGE ENVIRONMENTS 157

Requirements Assessment	158
Total Disk Space Requirements	159
Total Processing Power	161
Parallelization versus Serialization	164
Operating System Requirements	164
High Availability	165
I/O Bandwidth Requirements	166
Features	166
End-User Support	167
Sizing Machine Capacity	167
General Resource Planning	168
Assessing Resource Utilization	169
Building a Distributed Model	170
Round-Robin Distributed Networking	170
Distributed BGP Networking	172
Final Thoughts	174

PART III ADVANCED CONCEPTS OF STATISTICAL FILTERING

10 TESTING THEORY 177

The Challenge of Testing	178
Message Continuity	178
Archive Window	179
Purge Simulation	180
Interleave	181
Corrective Training Delay	181

Types of Simulations	181
Measuring the Accuracy of a Specific Filter	182
Test Criteria	182
Performing the Test	183
Measuring Adaptation in Chaotic Environments	185
Test Criteria	185
Performing the Test	186
Testing the Effectiveness of Multiple Filters	187
Test Criteria	188
Performing the Test	189
Comparing Features in a Single Filter	191
Test Criteria	191
Performing the Test	192
Testing Caveats	193
Corrective Training	193
Purge Simulations	194
Test Messages	194
Presuppositions	195
Final Thoughts	195

11

CONCEPT IDENTIFICATION: ADVANCED TOKENIZATION

197

Chained Tokens	198
Case Study Analysis	199
Pattern Identification	200
Differentiation	201
HTML Classification	202
Contextual Analysis	203
Other Uses	204
Administrative Concerns	205
Supporting Data	206
Summary	207
Sparse Binary Polynomial Hashing	207
Supporting Data	209
Summary	210
Karnaugh Mapping	210
Final Thoughts	213

12

FIFTH-ORDER MARKOVIAN DISCRIMINATION

215

Markov's Great Advance	216
Hidden Markov Models (HMMs)	218
Using Markov Models to Model Text	219
Classic Bayesian Spam Filter	219
Bayesian versus Markovian Classification	222
Storage Concerns	225
Purging Old Data	226

Floating-Point Renormalization and Underflow	226
Final Thoughts	226

13 INTELLIGENT FEATURE SET REDUCTION 227

Calibration Algorithms	228
Bayesian Noise Reduction (BNR)	231
Instantiation Phase	232
Training Phase	233
Dubbing Phase	234
Examples	236
End Result	239
Efficacy	239
Final Thoughts	240

14 COLLABORATIVE ALGORITHMS 241

Message Inoculation	242
Supporting Data	246
External Inoculation	246
Classification Groups	247
Collaborative Neural Meshes	248
Neural Declustering	249
Machine-Automated Blacklists	250
Streamlined Blackhole List	251
Weighted Private Block List	252
Distributed Attacks	252
Filters That Fight Back	252
Fingerprinting	253
Probing	253
Automatic Whitelisting	253
URL Blacklisting	255
Minefields	256
Final Thoughts	256

APPENDIX SHINING EXAMPLES OF FILTERING 257

POPFile: The POP3 Proxy	258
About POPFile	258
Accuracy	259
Interview with the Author	260
SpamProbe: A Modified Approach	261
About SpamProbe	261
Accuracy	262
Interview with the Author	262

TarProxy: IANA Spam Filter	264
About TarProxy	264
Accuracy	264
Interview with the Author	265
DSPAM: A Large-Scale Filter	266
About DSPAM	266
Accuracy	267
Interview with the Author	268
The CRM114 Discriminator	270
About CRM114	270
Under the Hood	271
Accuracy	272
Interview with the Author	272

INDEX

275