

Spis treści

Wprowadzenie	15
Od tłumacza	18
I Wstęp do filtrowania spamu	21
1 Historia spamu	23
1.1. Definicja spamu	23
1.2. Pierwszy spam	24
1.3. Spam: początki	27
1.3.1. Fundusz edukacyjny Jay-Jaya	28
1.3.2. Przesłanie religijne	30
1.3.3. Canter i Siegel	31
1.3.4. Cancelmoose	34
1.3.5. Jeff „Spam King” Slaton	36
1.3.6. Kevin „Krazy” Lipsitz	37
1.3.7. Stanford Wallace i Cyber Promotions	37
1.3.8. Floodgate — pierwsze oprogramowanie spamujące	38
1.3.9. Inne znaczące wydarzenia w 1995 roku	38
1.4. Spamowy wyścig zbrojeń	39
1.4.1. Spamhaus	39
1.4.2. Niezamawiana poczta reklamowa	41
1.5. Spam poza kontrolą	42
1.5.1. Lata 1998–2000: trzy lata wojny ze spamem	43
1.5.2. Network Solutions	44
1.5.3. Po 2001 roku: wykładniczy wzrost spamu	45
1.6. Rozważania końcowe	45

2	Historyczne podejścia do zwalczania spamu	47
2.1.	Prymitywna analiza językowa	47
2.2.	Czarne listy	49
2.2.1.	Problemy z propagowaniem i aktualizowaniem	50
2.3.	Filtrowanie heurystyczne	51
2.3.1.	Brightmail	52
2.3.2.	SpamAssassin	53
2.3.3.	Słabości filtrowania heurystycznego	53
2.3.4.	Poważne problemy z aktualizowaniem	54
2.3.5.	Ocenianie	54
2.4.	Białe listy	55
2.4.1.	Za mała efektywność	55
2.4.2.	Fałszerstwa	56
2.5.	Wezwanie-odpowiedź	57
2.5.1.	Problemy schematu wezwanie-odpowiedź	58
2.6.	Ograniczanie	58
2.6.1.	TarProxy	59
2.6.2.	Inne narzędzia ograniczające	60
2.7.	Filtrowanie zbiorowe	61
2.8.	Gmatwanie adresów	63
2.9.	Nowe standardy	64
2.9.1.	Uwierzytelnianie SMTP	64
2.9.2.	Założenia systemowe dotyczące wysyłania	65
2.10.	Kroki prawne	66
2.10.1.	Ślady pozostawiane przez natrętów	69
2.10.2.	Własność intelektualna	70
2.11.	Rozważania końcowe	71
3	Pojęcia związane z klasyfikacją językową	73
3.1.	Dokładność	74
3.2.	Nauczanie zautomatyzowane	74
3.2.1.	Nauczanie koncepcyjne	75
3.3.	Korzystanie z klasyfikacji językowej do zwalczania spamu	76
3.3.1.	Szkolenie	76

3.3.2.	Filtrowanie statystyczne i bayesowskie	77
3.4.	Składniki klasyfikatora językowego	78
3.4.1.	Dane historyczne	78
3.4.2.	Leksemizator	81
3.4.3.	Motor analityczny	82
3.5.	Informacje zwrotne	84
3.6.	Szkolenie	85
3.6.1.	Szkolenie pełne (TEFT)	85
3.6.2.	Szkolenie na podstawie błędów (TOE)	86
3.6.3.	Szkolenie na podstawie nowości (TUM)	87
3.6.4.	Szkolenie aż do wyeliminowania błędów (TUNE)	87
3.6.5.	Przeprowadzanie szkoleń	88
3.7.	Przykład filtra	88
3.7.1.	Etap 1. Leksemizacja wiadomości	89
3.7.2.	Etap 2. Konstruowanie macierzy decyzyjnej	90
3.7.3.	Etap 3. Wartościowanie macierzy decyzyjnej	91
3.7.4.	Etap 4. Używanie wiadomości w procesie szkolenia	91
3.7.5.	Etap 5. Korygowanie błędów	91
3.8.	Skuteczność filtrowania statystycznego	91
3.9.	Przyszłość klasyfikacji językowej	92
3.9.1.	Panowanie filtrowania statystycznego	93
3.10.	Rozważania końcowe	94
4	Podstawy filtrowania statystycznego	95
4.1.	Niedoskonałość rozwiązania	96
4.2.	Konstruowanie zbioru danych historycznych	96
4.2.1.	Analiza treści wiadomości	97
4.2.2.	Rozpoczynanie od pustego zbioru	98
4.2.3.	Korygowanie błędów	99
4.2.4.	Leksemizator i obliczanie wartości leksemów	100
4.2.5.	Leksemy kategoriyczne	103
4.2.6.	Filtry stroniczne	103
4.2.7.	Leksemy jednostkowe	104
4.2.8.	Wyniki końcowe	105

4.3.	Motor analityczny	105
4.3.1.	Sortowanie	107
4.4.	Kombinacje statystyczne	108
4.4.1.	Kombinacje bayesowskie (Paul Graham)	109
4.4.2.	Kombinacje bayesowskie (Brian Burton)	110
4.4.3.	Test średnich geometrycznych Robinsona	112
4.4.4.	Odwrotny test chi-kwadrat Fishera-Robinsona	113
4.5.	Udoskonalanie analizy statystycznej	114
4.5.1.	Udoskonalanie macierzy decyzyjnej	115
4.5.2.	Udoskonalanie leksemizacji	115
4.5.3.	Równoważenie statystyczne	116
4.5.4.	Szkolenie iteracyjne	117
4.5.5.	Poznanwanie nowych sztuczek	118
4.6.	Rozważania końcowe	119
II	Filtrowanie statystyczne	121
5	Dekodowanie wiadomości	123
5.1.	Wprowadzenie do kodowania	123
5.1.1.	Dekodowanie	124
5.2.	Kodowanie treści wiadomości	125
5.2.1.	Kodowanie znaków niedrukowalnych	127
5.2.2.	Kodowanie Base64	128
5.2.3.	Kodowanie niestandardowe	129
5.3.	Kodowanie nagłówek wiadomości	129
5.4.	Kodowanie HTML-owe	130
5.5.	Dekomponowanie wiadomości	132
5.6.	Oprogramowanie pomocnicze	132
5.7.	Rozważania końcowe	133
6	Leksemizacja: cegiełki spamu	135
6.1.	Leksemizacja jako funkcja heurystyczna	135
6.2.	Podstawowe ograniczniki	136
6.3.	Nadmiarowość	137
6.4.	Inne ograniczniki	138
6.5.	Wyjątki	139

6.6.	Zestawianie leksemów	140
6.7.	Degeneracja	140
6.8.	Optymalizowanie nagłówków	141
6.9.	Optymalizowanie adresów URL	143
6.10.	Leksemizacja HTML-a	144
6.11.	Pary wyrazów	147
6.12.	Haszowanie za pomocą rzadkich wielomianów binarnych	147
6.13.	Wielojęzyczność	148
6.14.	Rozważania końcowe	148
7	Nieczne sztuczki natrętów	151
7.1.	Skuteczność filtrowania	152
7.1.1.	Koniec z bólem głowy	152
7.2.	Słabości filtrowania statystycznego	153
7.3.	Ataki na leksemizatory	153
7.3.1.	Niewłaściwe kodowanie	154
7.3.2.	Kodowanie nagłówków	155
7.3.3.	Hipertekst przerywany	156
7.3.4.	Spam ASCII	158
7.3.5.	Rozdzielanie tekstu	160
7.3.6.	Gmatwanie tabelowe	162
7.3.7.	Kodowanie adresów URL	164
7.3.8.	Tekst symboliczny	166
7.3.9.	Inne głupstwa	166
7.4.	Ataki na zbiory danych	167
7.4.1.	Ataki na listy adresowe	168
7.4.2.	Zatruwanie bayesowskie	169
7.4.3.	Pozornie puste listy	173
7.5.	Ataki na macierze decyzyjne	175
7.5.1.	Spam obrazkowy	175
7.5.2.	Napisy losowe	177
7.5.3.	Domieszane wyrazy	179
7.5.4.	Ataki ukierunkowane	181
7.6.	Rozważania końcowe	183

8	Przechowywanie olbrzymich ilości danych	185
8.1.	Rozważania na temat pamięci zewnętrznej	185
8.1.1.	Miejsce na dysku	186
8.1.2.	Szybkość	186
8.1.3.	Blokowanie	187
8.1.4.	Przenoszalność	187
8.1.5.	Stanowość	187
8.1.6.	Przywracanie	188
8.1.7.	Opóźnienia operacji wejścia-wyjścia	188
8.1.8.	Dostęp bezpośredni	189
8.1.9.	Łatwość używania	189
8.2.	Struktura szkieletowa pamięci zewnętrznej	189
8.3.	Inne rozwiązania pamięci zewnętrznej	192
8.3.1.	Bezstanowe implementacje baz danych	192
8.3.2.	Stanowe rozwiązania SQL-owe	195
8.3.3.	Biblioteka PBL ISAM Petera Grafa	197
8.3.4.	SQLite	200
8.4.	Implementacje własne	202
8.5.	Rozważania końcowe	202
9	Skalowanie w olbrzymich środowiskach	205
9.1.	Szacowanie wymagań	206
9.1.1.	Wymagania dotyczące łącznego miejsca na dyskach	207
9.1.2.	Łączna moc obliczeniowa	210
9.1.3.	Wykonywanie równoległe a sekwencyjne	213
9.1.4.	Wymagania dotyczące systemów operacyjnych	214
9.1.5.	Gwarancja nieprzerwanego działania	214
9.1.6.	Wymagania dotyczące przepustowości systemu wejścia-wyjścia	215
9.1.7.	Funkcjonalność	216
9.1.8.	Obsługa użytkowników końcowych	217
9.2.	Określanie możliwości komputerów	217
9.2.1.	Ogólne planowanie zasobów	218
9.2.2.	Szacowanie wykorzystywania zasobów	219

9.3.	Konstruowanie modelu rozproszonego	221
9.3.1.	Karuzelowe rozproszone przetwarzanie sieciowe	221
9.3.2.	Bramkowe rozproszone przetwarzanie sieciowe	223
9.4.	Rozważania końcowe	225
III	Zaawansowane pojęcia związane z filtrowaniem statystycznym	227
10	Teoria testowania	229
10.1.	Wyzwanie testowania	229
10.1.1.	Ciągłość wiadomości	230
10.1.2.	Okresy archiwizowania	232
10.1.3.	Symulowanie kasowania	232
10.1.4.	Przeplatanie	233
10.1.5.	Opóźnienia szkoleń korygujących	233
10.2.	Typy symulacji	234
10.3.	Mierzenie dokładności konkretnych filtrów	235
10.3.1.	Kryteria testowe	235
10.3.2.	Przeprowadzanie testów	236
10.4.	Mierzenie adaptowania się w środowisku chaotycznym	238
10.4.1.	Kryteria testowe	239
10.4.2.	Przeprowadzanie testów	240
10.5.	Testowanie efektywności wielu filtrów	241
10.5.1.	Kryteria testowe	242
10.5.2.	Przeprowadzanie testów	243
10.6.	Porównywanie funkcji pojedynczego filtra	245
10.6.1.	Kryteria testowe	246
10.6.2.	Przeprowadzanie testów	247
10.7.	Problemy z testami	248
10.7.1.	Szkolenia korygujące	248
10.7.2.	Symulowanie kasowania	249
10.7.3.	Wiadomości testowe	250
10.7.4.	Powody problemów	250
10.8.	Rozważania końcowe	251

11 Identyfikacja pojęciowa: zaawansowana analiza leksykalna	253
11.1. Leksemy łańcuchowe	254
11.1.1. Analizy studiów przypadków	255
11.1.2. Identyfikowanie wzorców	256
11.1.3. Rozróżnianie	258
11.1.4. Klasyfikowanie HTML-a	259
11.1.5. Analiza kontekstowa	260
11.1.6. Inne przypadki	261
11.1.7. Zagadnienia administracyjne	262
11.1.8. Dane pomocnicze	263
11.1.9. Podsumowanie	264
11.2. Haszowanie za pomocą rzadkich wielomianów binarnych	265
11.2.1. Dane pomocnicze	267
11.2.2. Podsumowanie	268
11.3. Odwzorowanie Karnaugha	268
11.4. Rozważania końcowe	271
12 Rozróżnienia markowskie piątego rzędu	273
12.1. Zastosowania modeli Markowa	274
12.2. Ukryte modele Markowa	276
12.3. Korzystanie z modeli Markowa do modelowania tekstu	277
12.3.1. Klasyczne bayesowskie filtry spamu	278
12.4. Klasyfikacja bayesowska a markowska	282
12.5. Zagadnienia związane z pamięcią zewnętrzną	285
12.5.1. Kasowanie przestarzałych danych	286
12.6. Renormalizacja i niedomiary zmiennopozycyjny	286
12.7. Rozważania końcowe	287
13 Inteligentne redukowanie zbioru funkcji	289
13.1. Algorytmy kalibracji	290
13.2. Bayesowska redukcja szumów	294
13.2.1. Faza tworzenia instancji	295
13.2.2. Faza szkolenia	296
13.2.3. Faza podstawiania	298

13.2.4.	Przykłady	299
13.2.5.	Wynik końcowy	303
13.2.6.	Skuteczność	303
13.3.	Rozważania końcowe	304
14	Algorytmy zbiorowe	307
14.1.	Wiadomości jako szczepionki	307
14.1.1.	Dane pomocnicze	312
14.1.2.	Szczepienia zewnętrzne	312
14.2.	Grupy klasyfikacyjne	313
14.3.	Zbiorowe sieci neuronowe	314
14.3.1.	Rozdrabnianie neuronowe	316
14.4.	Zautomatyzowane czarne listy	318
14.4.1.	Uproszczone listy blokujące	318
14.4.2.	Prywatne ważne listy blokujące	319
14.4.3.	Ataki rozproszone	319
14.5.	Filtry kontratakujące	320
14.6.	Tropienie śladów	320
14.7.	Kontrolowanie serwera	321
14.8.	Zautomatyzowane białe listy	321
14.9.	Czarne listy adresów URL	323
14.10.	Pola minowe	324
14.11.	Rozważania końcowe	325
 Dodatek		
Przykłady skutecznego filtrowania		
327		
POPFile: Proxy POP3		327
POPFile — opis		327
Dokładność		329
Rozmowa z twórcą filtra		330
SpamProbe: Zmodyfikowane podejście		331
SpamProbe — opis		331
Dokładność		332
Rozmowa z twórcą filtra		333

TarProxy: Filtr spamu IANA	334
TarProxy — opis	335
Dokładność	335
Rozmowa z twórcą filtra	336
DSPAM: Filtr wielkoskalowy	338
DSPAM — opis	338
Dokładność	339
Rozmowa z twórcą filtra	340
CRM114 Discriminator	342
CRM114 — opis	343
Budowa wewnętrzna	343
Dokładność	344
Rozmowa z twórcą filtra	345
Skorowidz	349