ジャンクメールフィルタ入門

Mozilla Junk Mail Control

Mozilla dot Party in Japan 4.0

2003年4月19日

えむもじら　江村 秀之

http://www5e.biglobe.ne.jp/~level0/mozilla/

目次

- スパム (ジャンクメール) とは
- スパムフィルタ
- ベイジアン手法
- Mozillaのジャンクメールコントロール
- デモ
- まとめ

スパム（ジャンクメール）とは

- スパムの定義
  - 受け手の意に反して配信される電子メール
    （広告、ウィルス、チェーンメールなど）
- スパムの問題点
  - 受け手にコストが発生する（通信費、手間など）
  - ネットーワークやサーバーに負荷がかかる
- 郵便との違い
  - 差出人のコストが圧倒的に少ない
  - 1000通当たりのコスト(GartnerG2)
    郵便：$500～$700、電子メール：$5～$7

スパムの流通量

- **私の場合：3～4通/日**
- AOL (2003/2/21:ZDNNニュース)
  - … 同社は最近、約2700万人の米国会員から、…
    発表によると同社の独自スパム排除技術は毎日7億8000万件のジャンクメールを遮断している。これは会員1人当たりで言うと、1日22件のスパムが遮断されていることになる。
- Symantec社 (2002/12/6:BizTech)
  - 職場や家庭で1週間当たりに「100通以上」のスパム・メールを受信するユーザーは37%、「50通以上」受信するユーザーは63%に達したという。
- Brightmail社 (2002/8/30:ZDNNニュース)
  - 2002年7月には、インターネット上でやり取りされる電子メールの36%をスパムが占めた。約1年前は8%だった。
- David Mertz, Ph.D. (「スパムの選り分け手法」著者)
  - … 毎日、正当なメールのやりとりの *何倍も* 多くのスパムをを受けとっています。平均して、多分、正当なEメール1個に対して10個のスパムを受けとっています。

スパム排除手法（スパムフィルタ）

- 基本的な構造化テキスト・フィルター
- ホワイトリスト＋自動照合
  - 信頼できる相手からのみ受信
  - 未知の相手には確認メールを要求
- 分散適応型ブラックリスト
  - サーバーのスパム情報にアクセス
- ルールベース
- ベイズ単語分布フィルター
- ベイズの三連文字フィルター

文献[3]より

ベイズ理論

- ベイズ理論
  - 「未来を推測するには過去を振り返らなければならない」
  - ベイズ理論では、完全に現実の世界から集められたデータに基づいて推測を行い、データの数が多ければ多いほどより確実な推測を引き出せる。また、ベイズモデルは自己修正型モデルであり、データの変化に応じて結果が変わる。

文献[4]より

ベイズフィルタ

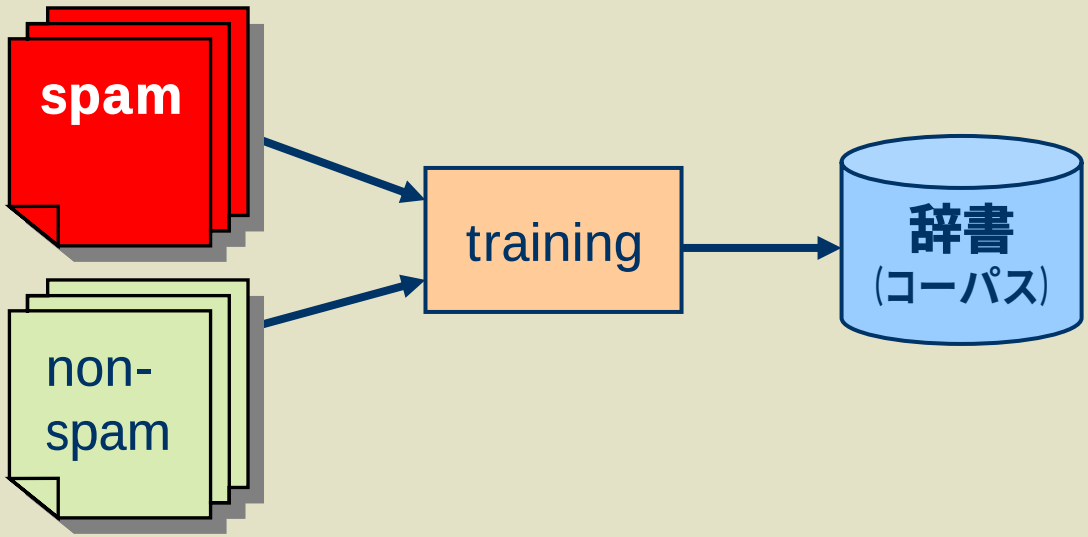
- ベイズ理論に基づくメールフィルタ
  - 「特定の単語はスパムに高頻度で出現し、別の単語は非スパムに高頻度で出現する」
  - まず、スパムと非スパムから辞書（コーパス）を作成（学習）
  - 単語単位にスパム確率を計算。
  - メールに含まれる単語のスパム確率から、そのメールのスパム確率を計算

単語の抽出

- 英語
  - 単語：適当な区切り文字で分離
  - 3連文字：3文字単位で抽出
- 日本語
  - 単語：抽出難しい。
    精度あげるには構文解析、辞書必要。
  - 2連文字 (bigram) ：辞書不要
    例：「裏ビデオ販売」
    →「裏ビ」、「ビデ」、「デオ」、「オ販」、「販売」

学習データから抽出する情報

- 学習したスパム数(nbad)、非スパム数(ngood)
- 単語およびその単語が
  - スパムに表れた回数(b)
  - 非スパムに表れた回数(g)
  - その単語のスパム確率 (上記から計算)



単語のスパム確率

単語のスパム確率 =

$$\begin{cases} \dfrac{\min(1.0,\ b/nbad)}{\min(1.0,\ 2g/ngood) + \min(1.0,\ b/nbad)} & (2g+b>5) \\ 0.4 & (others) \end{cases}$$

ただし、0.01を下限、0.99を上限とする。

メールのスパム確率

メールのスパム確率 =

$$\frac{p1*p2*\ldots*p15}{p1*p2*\ldots*p15 + (1-p1)*(1-p2)*\ldots*(1-p15)}$$

p*n* はメール中のもっとも特徴的な（0.5から最も離れている）単語15個のスパム確率

スパム確率>0.9のものをスパムとする。

ベイズ手法の利点

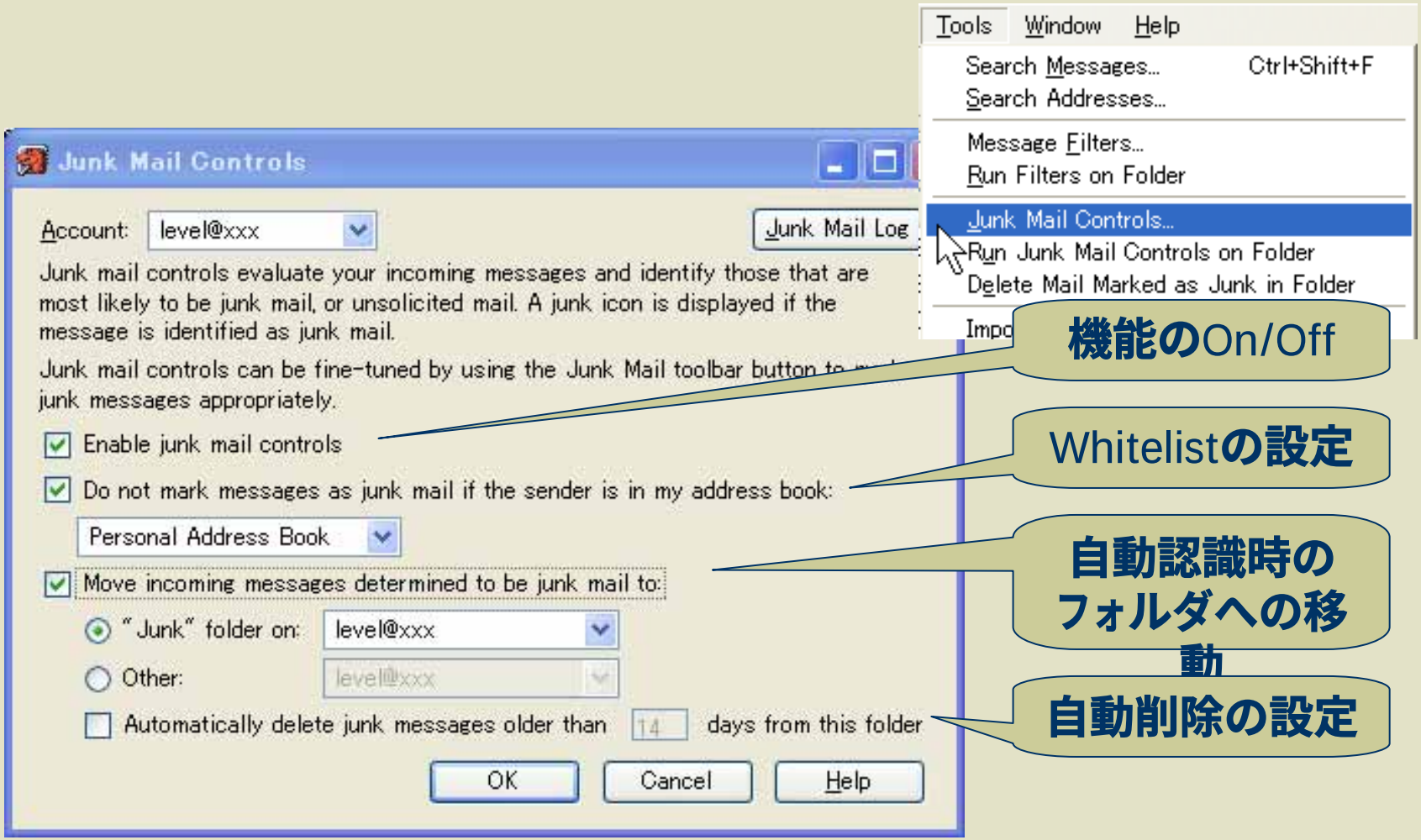
- ルールの作成が容易
  - 分類されたメールから自動的に作成
- 学習結果が個人ごとに異なる
  - 個人毎に異なるスパム定義
  - フィルタを回避するのが困難
- メンテナンスが不要
  - 新しい種類のスパムも自動的に学習してくれる
- アルゴリズムが単純
  - 処理が高速
- 非常に高精度

Mozillaのジャンクメールコントロール

- Mozilla 1.3から実装
- 手順
  1. ジャンクメールコントロールの有効化
  2. トレーニング
     - 実際のメールでspamとnon-spamを学習させる
  3. メール受信
  4. 間違った判定を修正 --- 重要
     - ジャンクアイコン
     - ジャンクバー

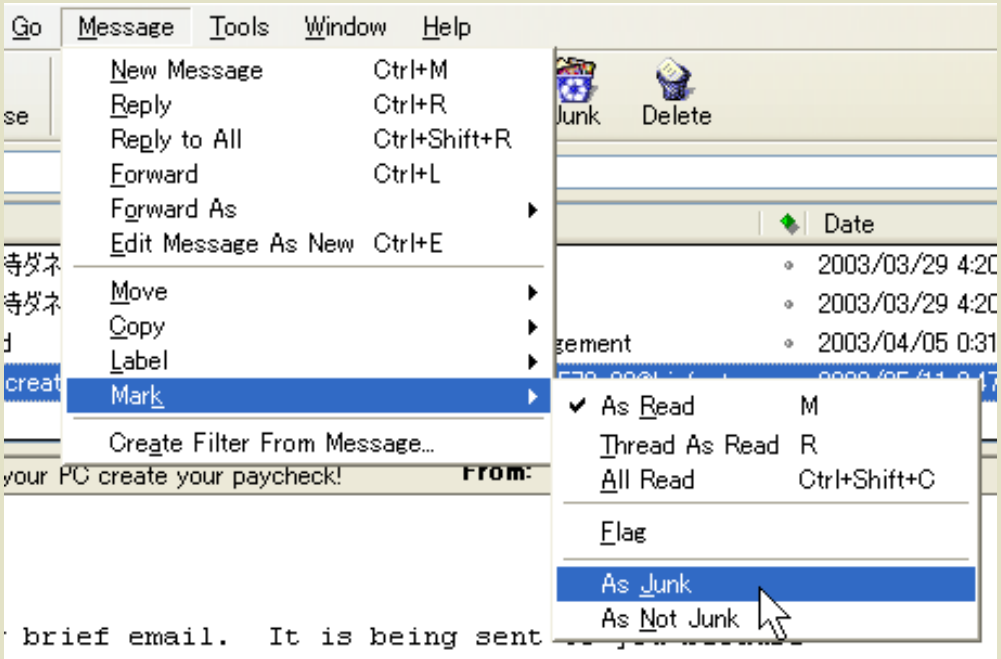
Mozilla 1.4a

ジャンクメールコントロール画面



トレーニング

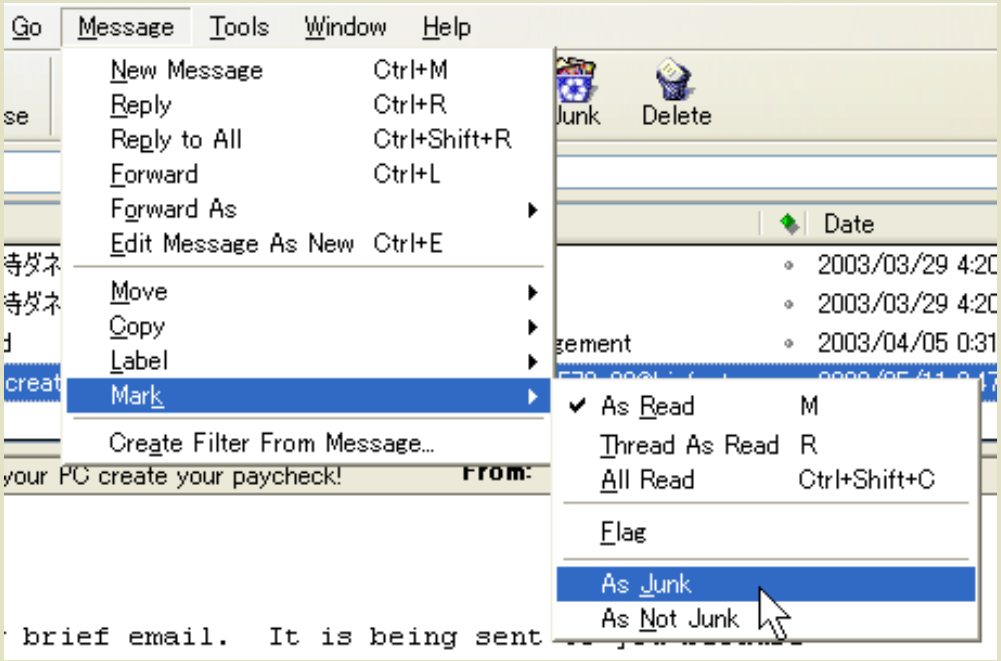
- まず、スパムと非スパムの集合を用いてトレーニング (学習) を行う必要がある。
  - Message->Mark->As Junk
  - Message->Mark->As Not Junk
  - Junkボタン



メールウィンドウ

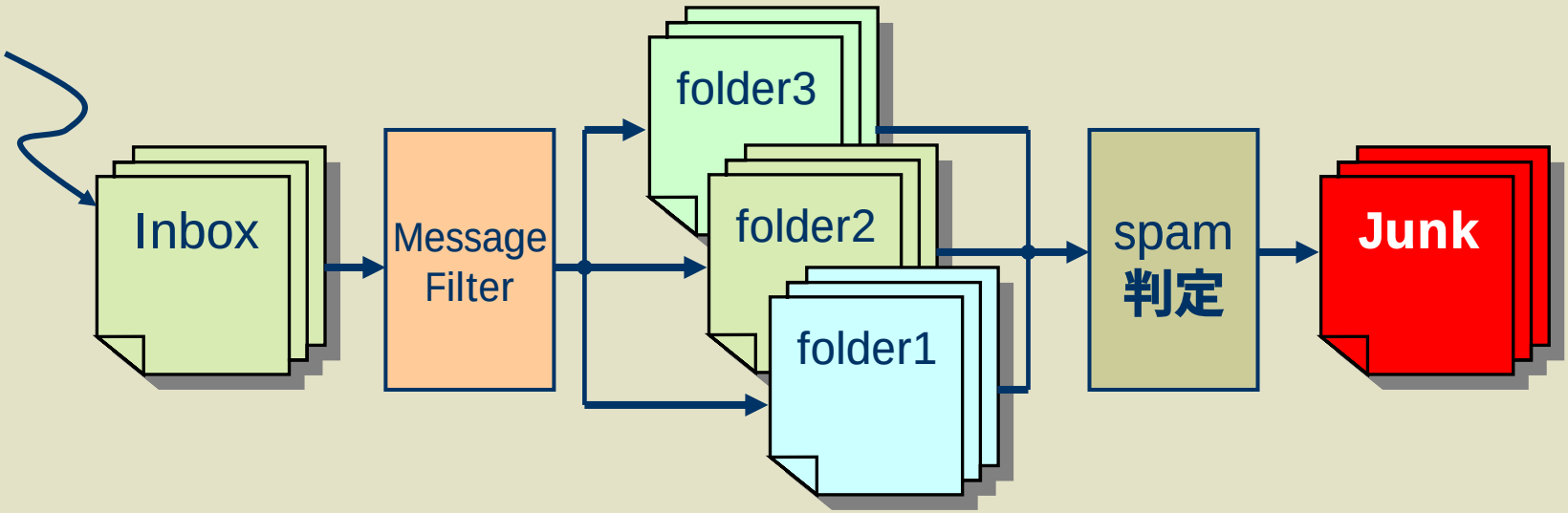

Junk Icon
Junk Bar
Junkフォルダ
Signs from God - Junk for level@xxx - Mozilla
File Edit View Go Message Tools Window Help
Get Msgs Compose Reply Reply All Forward Next Not Junk Delete
Name
level@xxx
Inbox
Drafts
Templates
Sent
Junk
Trash
JunkTest
View: All
Subject or Sender contains:
Subject Sender Priority
未承諾広告※特ダネ情報 me464783 2003/03/29 ...
Signs from God The last Judgement 2003/04/05 ...
Mozilla thinks this message is junk mail
?
Not Junk
Subject: Signs from Goc From: The last Judgement 2003/04/05 0:31
Signs from God. The Messiah comes. We have the end of the World
Signs from God. The Messiah comes. We have the end of the World and already 3th
World war. The Mankind faces the Doom and as well the biggest ever experienced
Unread: 0
Total: 254

- まず、スパムと非スパムの集合を用いてトレーニング (学習) を行う必要がある。
  - Message->Mark->As Junk
  - Message->Mark->As Not Junk
  - Junkボタン



メール受信時のspam認識の流れ

1. メール受信
2. Message Filterによる振り分け
3. 受信メッセージに対するspam判定
4. Junkフォルダへの移動



辞書ファイル：training.dat

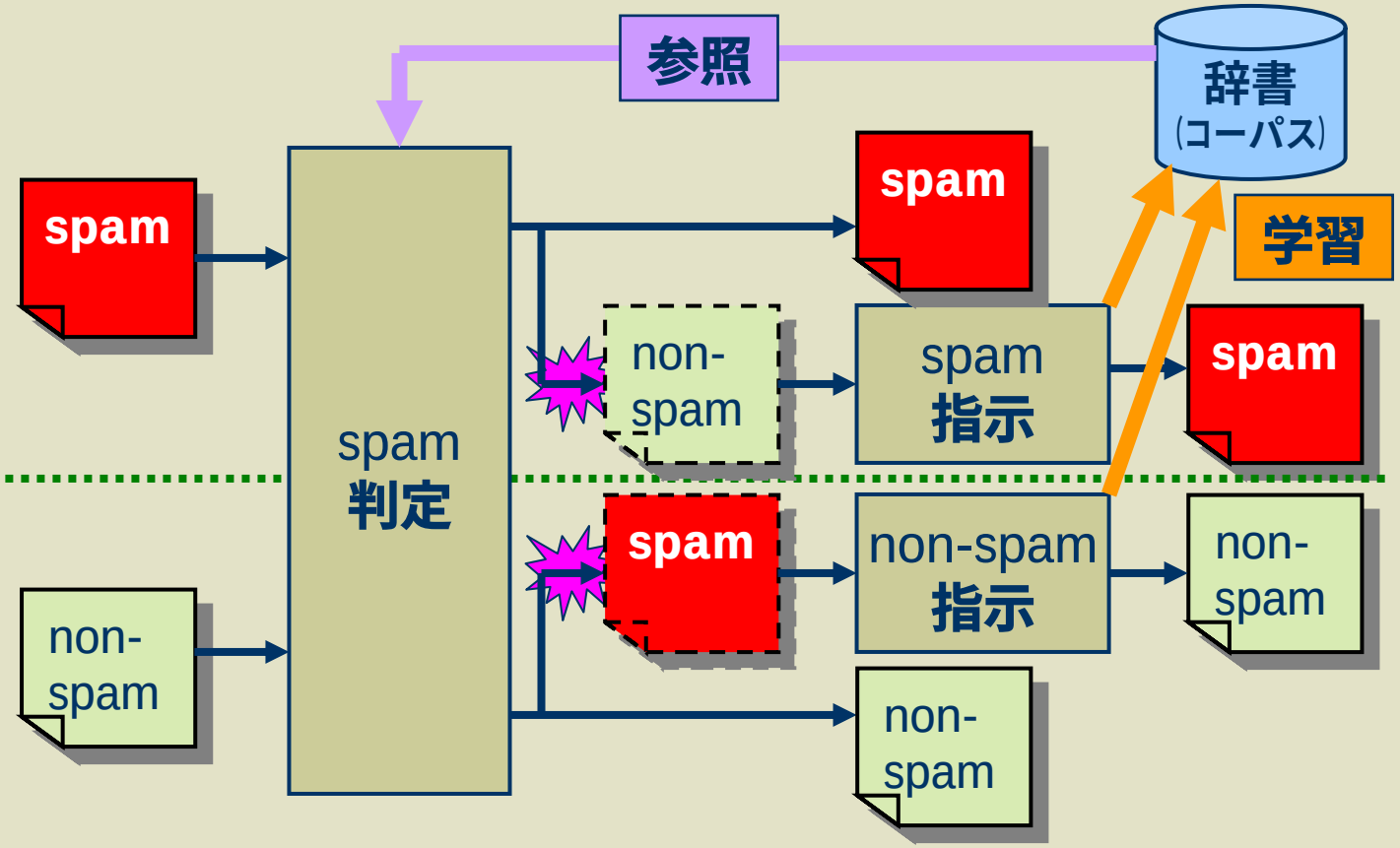
- ジャンクメールコントロールにおける辞書情報（コーパス）を格納
- 場所：プロファイルフォルダ
- 全アカウント共通

学習する対象

- ヘッダも含めたメール全体
  - 対象になるもの：
    - ヘッダ、本文、メールアドレス、サーバーアドレス、HTMLのタグ、等
  - 対象外：
    - バイナリの添付ファイル

学習するタイミング

- 受信時：学習しない
- スパム/非スパムを明示的に指示：学習する



training.datの構造

	項目	サイズ	値
Header	cookie	4byte	FE ED FA CE
	GoodCount	4byte	non-spam文書数
	BadCount	4byte	spam文書数
non-spam情報	tokenCount	4byte	単語数
	count	4byte	単語の出現回数
	size	4byte	単語の長さ(バイト数)
	string	size byte	単語(UTF-8)
spam情報	tokenCount	4byte	単語数
	count	4byte	単語の出現回数
	size	4byte	単語の長さ(バイト数)
	string	size byte	単語(UTF-8)

mailnews/extensions/bayesian-spam-filter/src/nsBayesianFilter.cpp#965
nsBayesianFilter::readTrainingData()

Mozillaの単語処理

- 日本語
  - 連続した「ひらがな」
  - 連続した「カタカナ」
  - 連続した「記号」
  - 1文字の「漢字」

training.datの中身

スパムメール

いつもクラップメールご利用」ありがとうございます。

今日の紹介は

日本一の割り切った交際掲示板　使い放題

http://abc.co.jp/xxx/index.html

もし会員でない方や　配信を希望されない方は

　お手数ですがterurumu@abc.co.jpにメールアドレスのみを書いて

メールください。

training.datの単語例

```
1 いつも
1 クラップメール
1 ご
1 利
1 用
1 」
1 ありがとうございます
2 。
1 今
2 日
3 の
1 紹
1 介
2 は
1 本
1 一
```

単語処理の改善案

- 日本語
  - 2文字以上連続した「ひらがな」
  - 2文字以上連続した「カタカナ」
  - 2文字以上連続した「記号」
  - 1文字の「漢字」+任意の1文字
  - 2文字以上の連続した「漢字」

ただし、根拠なし！

改善案の例

スパムメール

いつもクラップメールご利用」ありがとうございます。

今日の紹介は

日本一の割り切った交際掲示板　使い放題

http://abc.co.jp/xxx/index.html

もし会員でない方や　配信を希望されない方は

　お手数ですがterurumu@abc.co.jpにメールアドレスのみを書いて

メールください。

単語例

いつも
クラップメール
ご利
利用
用」
ありがとうございます
今日
日の
の紹
紹介
介は
日本
本一
一の
の割
割り

training.datをダンプする

- **以下でPerlスクリプトを配布中**
  http://www5e.biglobe.ne.jp/~level0/mozilla/spam/

```
pmin=0.000000 pmax=1.000000 bmin=0
nGood 1305                              #学習した非スパム数
nBad    427                             #学習したスパム数
total non-spam tokens 36400             #非スパムに現れた単語数
total spam tokens 17725                 #スパムに現れた単語数
1 4 0.859 proven                        #数値の意味は
1 2 0.400 richer                        #1:非スパムに現れた数
1 0 0.400 mailto～              #2:スパムに現れた数
1 0 0.400 などがきちんとでき           #3:スパム確率
0 8 0.990 farms
1 0 0.400 でどうなるかわからない
6 0 0.010 bounces-to
1 0 0.400 そば
7 2993 0.989 nbsp
…
```

デモ

振り分け結果

	non-spam		spam		
判定	non-spam	spam	non-spam	spam	合計
4月10日	25	0	0	1	26
4月11日	24	0	0	4	28
4月12日	25	1	0	6	32
4月13日	16	0	0	1	17
4月14日	31	0	0	3	34
4月15日	35	0	1	2	38
4月16日	35	0	0	2	37
4月17日	28	0	0	1	29
4月18日	19	0	0	0	19
合計	238	1	1	20	260.0
	91.5%	0.4%	0.4%	7.7%	
平均	26.4	0.1	0.1	2.2	28.7

まとめ

- ベイズ手法によるMozillaのジャンクメールコントロール
- 簡単な操作、専門知識不要
  - ユーザ、管理者共
- 高い認識精度
- ダウンロードする手間は変わらず
  - 他のスパムフィルタとの併用
- すべてのユーザに、スパムフィルタを！

参考文献

- Paul Graham

  [1] A Plan for Spam（スパムへの対策）
  http://www.shiro.dreamhost.com/scheme/trans/spam-j.html

  [2] Better Bayesian Filtering（ベイジアンフィルタの改善）
  http://www.shiro.dreamhost.com/scheme/trans/better-j.html

- David Mertz

  [3] **スパムの選り分け手法**
  http://www-6.ibm.com/jp/developerworks/linux/021129/j_l-spamf.html

- CNET Japan

  [4] **グーグル、インテル、**MS**が注目するベイズ理論**
  http://japan.cnet.com/news/special/story/0,2000047679,20052855,00.htm

- mozilla.org

  [5] Mozilla **スパムフィルタ機能**
  http://jt.mozilla.gr.jp/mailnews/spam.html

  [6] Junk Mail Controls（UI**仕様**）
  http://www.mozilla.org/mailnews/specs/spam/