Bayesian Filteringを用いたspam対策

明治大学情報システム管理課

服部裕之
hhat@isc.meiji.ac.jp

MEIJI UNIVERSITY

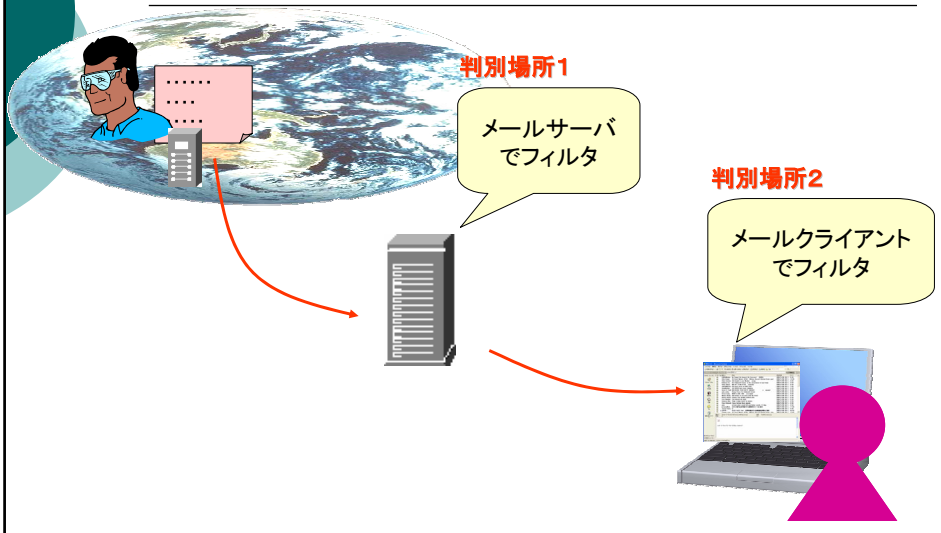
spamだらけの朝

朝の仕事は、まず、メールを消すことから

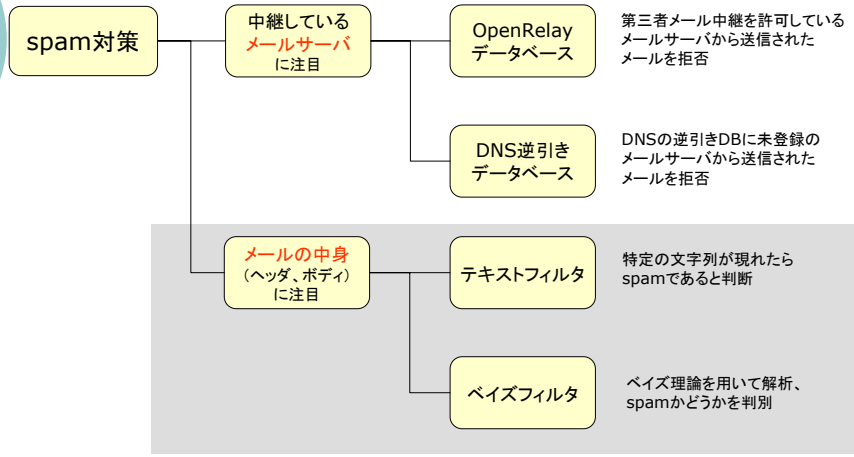
本日のメニュー

- spamを排除する方法は？
- 今、注目の「Bayesian Filtering」とは？
- 使いものになるのか？
- GraceMailからの利用は可能か？

spamの排除方法

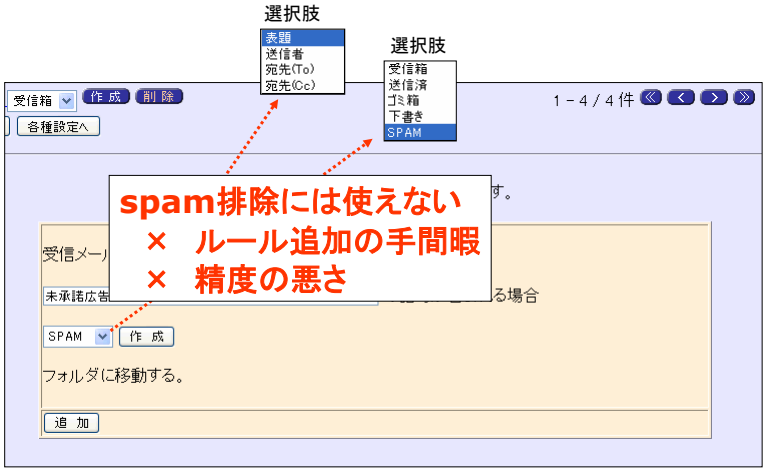


spam対策技術



テキストフィルタ

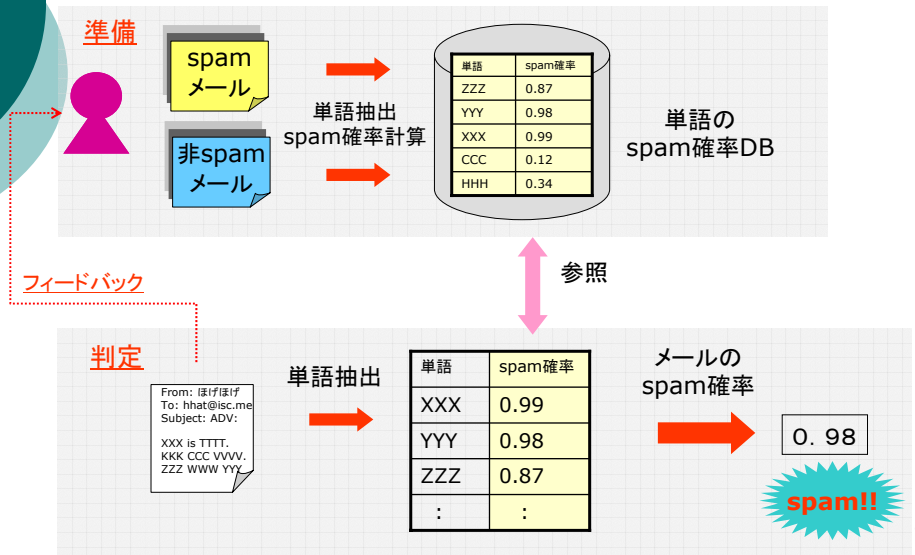
例： GraceMail 「メール振り分け機能」



ベイズフィルタ　- Bayesian Filtering -

- Bayes理論を応用したメールの分類手法
  - 前提
    - 「特定の単語」はspamに高頻度で出現する。
    - 「それ以外の単語」は非spamに高頻度で出現する。
  - よって、、、、
    1. 新規に到着したメールに含まれる単語を、過去のspam,非spamメールに含まれている単語と比較することによって、そのメールがspamであるかどうかを自動的に判別可能であるはず。
    2. 判別結果を、フィードバックすれば、より高精度の自動分類が可能になるはず。

Bayesian Filtering の仕組み



単語のspam確率の求め方
(Paul Graham式の場合)

単語のspam確率 =

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

非spam単語を重みづけ

出現頻度の低い単語は計算しない

ただし、0.01を下限、0.99を上限とする。

b	その単語がspamメール中に現れた回数
g	その単語が非spamメール中に現れた回数
nbad	spamメールの総数
ngood	非spamメールの総数

メールのspam確率の求め方
(Paul Graham式の場合)

メールのspam確率 =

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

pn はメール中の特徴的な単語(=0.5から最も離れている)15個のそれぞれのspam確率

メールのspam確率 > 0.9 を spamメールである、と判断

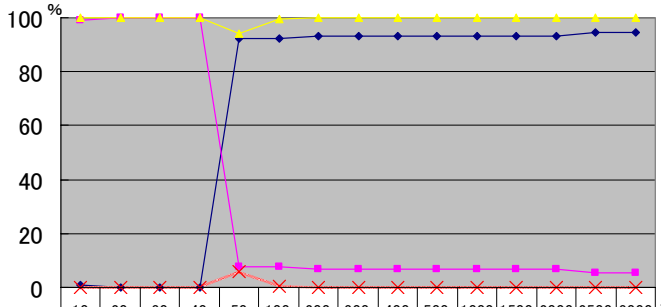
Bayesian Filtering の評価

- 使い物になるのか？
  - 判別精度はどのくらいあるのか？
  - 何通くらいのメールを学習すればよいのか？

評価に用いたプログラム

- bsfilter
  - Nabeya Kenichi氏作
  - GPL
  - rubyで記述
  - 3つの動作モード
    - フィルター
    - POPプロキシー
    - IMAPプロキシー
  - http://www.h2.dion.ne.jp/~nabeken/bsfilter/

spam確率DBの作成に使用したメール数と判別精度

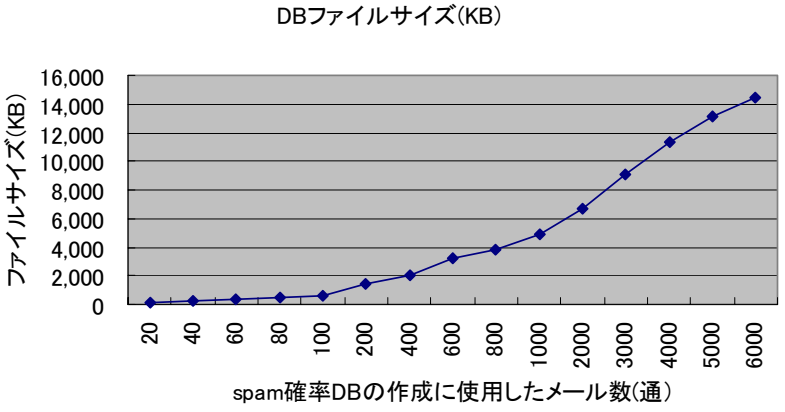


	10	20	30	40	50	100	200	300	400	500	1000	1500	2000	2500	3000
spamをspamと判定	1	0	0	0	92.5	92.5	93	93	93	93	93	93	93	94.5	94.5
spamを非spamと誤判定	99	100	100	100	7.5	7.5	7	7	7	7	7	7	7	5.5	5.5
非spamを非spamと判定	100	100	100	100	94	99.5	100	100	100	100	100	100	100	100	100
非spamをspamと誤判定	0	0	0	0	6	0.5	0	0	0	0	0	0	0	0	0

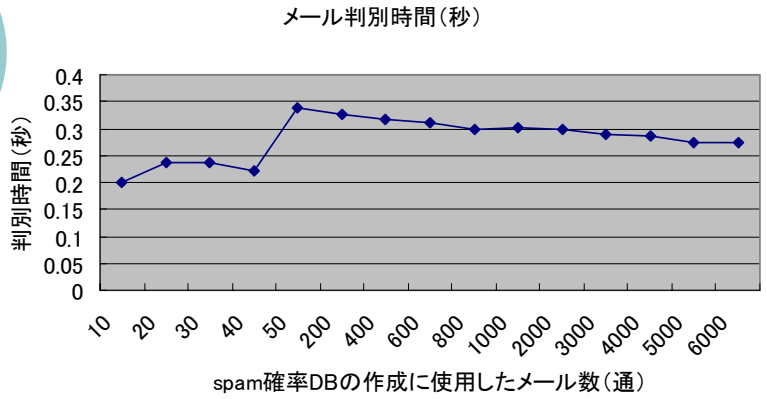
通

・**横軸**:spam確率DBを作成するために用いた、spam,非spamメール数 （通）
（12月7日までに hhat 宛に到着したメールを使用）
・**縦軸**:新着メールの判定精度 （%）
（12月8日以降に hhat 宛に到着したメールを使用。 spam、非spamそれぞれ200通を判別）

spam確率DBの作成に使用したメール数とspam確率DBのファイルサイズ

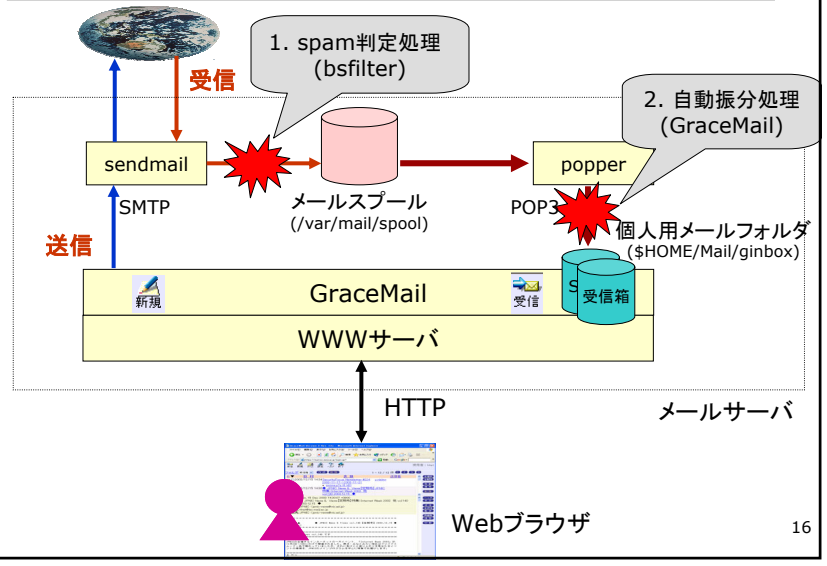


spam確率DBの作成に使用したメール数とメールの判別時間



実行環境:
SunBlade100 (UltraSPARCIIe 500MHz, mem:256MB, Solaris8)

GraceMail での Bayesian Filtering



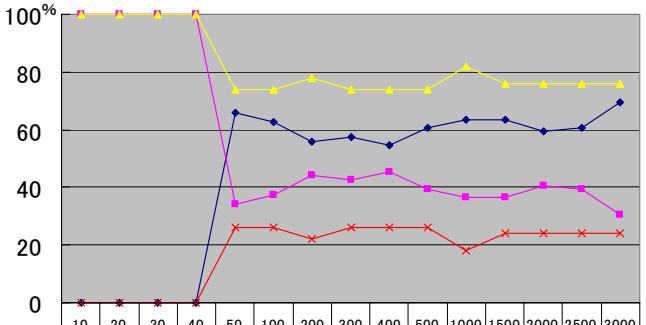
デモ

https://mail.kisc.meiji.ac.jp/

問題点

- 速度
  - spam単語データベースの更新に時間がかかる
- ディスク消費量
  - 学習用のメール
    - 4000通のメールで40MB程度
  - spam確率データベース
    - 4000通のメールから作成したデータベースのサイズは12MB程度

他人のspam確率DBを使って spamの判別は可能か？



	10	20	30	40	50	100	200	300	400	500	1000	1500	2000	2500	3000	通
spamをspamと判定	0	0	0	0	66	62.7	56	57.3	54.7	60.7	63.3	63.3	59.3	60.7	69.3	
spamを非spamと誤判定	100	100	100	100	34	37.3	44	42.7	45.3	39.3	36.7	36.7	40.7	39.3	30.7	
非spamを非spamと判定	100	100	100	100	74	74	78	74	74	74	82	76	76	76	76	
非spamをspamと誤判定	0	0	0	0	26	26	22	26	26	26	18	24	24	24	24	

・**横軸**：spam確率DBを作成するために用いた、spam,非spamメール数 （通）
（12月7日までに hhat 宛に到着したメールを使用）
・**縦軸**：新着メールの判定精度 （%）
（www-admin@meiji.ac.jp宛のメールを使用。 spam：150通、非spam：50通を判別）

spamを非spamと誤判別したメールの例

- すべて日本語で書かれたspamメール
- subjectに「未承諾広告」表示無し

Subject: インパクトのある卒業記念品のご提供

事務局ご担当者様
突然のご連絡大変失礼いたします。
総合通信販売を展開しております株式会社xxxxのxxxと申します。
今回ご連絡をさせて頂きましたのは、今までにない卒業記念品のご提供をさせて頂きたく、失礼とは承知の上、ご連絡をさせて頂きました。

- 中立を装った短いメール

Feel Younger.

find out about it here
http://www.happyhealthyfun.biz/

非spamをspamと誤判別したメールの例

- すべて入学に関する問い合わせメール
- 英語

Subject: Admission
To: www-admin@meiji.ac.jp

Dear Sir
I am student of Banglades.
I Completed my Bachelor Degree and Masters Degree in Education at Institute of Education, University of Dhaka, Bangladesh.
However, I want to get chance to admit to your university at M.Phil with P.hd. programme.
What can I do for? Plz. inform me.

Do you Yahoo!?
Protect your identity with Yahoo!
Mail AddressGuard http://antispam.yahoo.com/whatsnewfree

これはまずい！

まとめ

- Bayesian Filtering は使えるか？
  - Yes
  - 但し、他人宛のメールで学習したspam確率DBを使う時は要注意（www-adminの例）
- 何通くらい学習すれば使い物になるか？
  - 最低400通（spam,非spam各200通）
- GraceMailから使えるか？
  - Yes
  - 但し、ディスク消費量には要注意

参考文献

- Paul Graham
  - スパムへの対策 ---A Plan for Spam
    - http://www.shiro.dreamhost.com/scheme/trans/spam-j.html
  - ベイジアンフィルタの改善 --- Better Bayesian Filtering
    - http://www.shiro.dreamhost.com/scheme/trans/better-j.html
- Spam Detection
  - http://radio.weblogs.com/0101454/stories/2002/09/16/spamDetection.html
- spamメールのアーカイブ
  - http://www.spamarchive.org/
- spamassasin
  - http://useast.spamassassin.org/index.html

[参考1]　bsfilter使用例

- 単語のspam確率DBの更新
  - spamメールの単語を確率DBに追加
    bsfilter --update --spam < メール全文
  - 非spamメールの単語を確率DBに追加
    bsfilter --update --clean < メール全文
- 新着メールの判定
  - メールのspam確率を求める
    bsfilter < 新着メール
  - 判定結果をメールのヘッダに挿入
    bsfilter --insert-probability --insert-flag < 新着メール

[参考2]　GraceMailへの応用（設定例）

1. メールの転送設定
   - $HOME/.forward　ファイル

```
"| /home3/ob00020/bin/procmail"
```

2. procmailの設定
   - $HOME/.procmailrc　ファイル

```
:0 fw
|/home3/ob00020/bsfilter --pipe --insert-flag --insert-probability
:0 fw
* ^X-Spam-Probability: *(1|0¥.[6789])
|/home3/ob00020/bin/subject.pl
```

3. subject.pl
   - "Subject:タイトル"　→　"Subject: 未承諾広告※: タイトル"　に変換
4. GraceMail
   - 「設定」→「メール振分けルールの設定」
   - 題名に「未承諾広告※」が含まれれば、spamフォルダへ振り分け