bogofilter-sqlite クイックセットアップ

2017/02/19
Bogofilterの辞書が壊れて放置していたのを、やっと作り直すついでで、性能が改善していると聞いた Bogofilter-sqlite に切り替える。

2016/04/02 PostfixもSPFで検証ができるようにした時の備忘録

bogofilter-sqlite の導入

bogofilterよりメール文面内形態素解析の品質が良くなったという話を見たので、こちらを採用する。
やっぱりオリジナルは日本語に弱いみたいだし、少しだけ期待している。

bogofilter-sqlite のインストール

うちの環境は FreeBSD 11.0 RELEASE p1 で、bogofilter-sqlite は portsから導入。
cd /usr/ports/mail/bogofilter-sqlite; make; make install を実行。依存関係でエラー等あれば臨機応変に対応。

bogofilterとは同居できないので一旦アンインストールする。

pkg info コマンドを叩くと、パッケージ名は bogofilter-sqlite-1.2.4_2 になっていた。

```
$ pkg info | grep bogofilter
bogofilter-sqlite-1.2.4_2      Fast, teachable, learning spam detector
$
```

bogofilter-sqlite の設定

bogofilterもそうだがbogofilter-sqliteは事前に学習させる必要がある。

今回は

種別	件数
SPAM	159,921 件
NonSPAM	46,388 件
合計	206,309 件

のメールを学習させた。

……で、このドキュメント作成時点でSPAM学習がまだ終わっていない NonSPAM学習は約11時間で終わったSPAM/NonSPAMを並列で学習させたせいかもしれない。後ほど説明する。

wordlist.db の削除

この環境に残る、bogofilterで使っていた wordlist.db を消す。
bogofilter-sqliteが同名のsqlite用データベースファイルを作成する際に問題となる。

```
$ rm ~/.bogofilter/wordlist.db
$
```

SPAM/NonSPAMの学習

うちの環境はPostfixでMaildir形式のメールボックスなので、メールは1件1ファイルで格納されている。
これをコピーして1メール毎にSPAM/NonSPAMとして読み込ませ学習さる。

```
$ find /home/k896951/Maildir/.spam/cur -type f -print0 | xargs -0 -n 1 -P 5
-I{} bogofilter -svI '{}'
$ find /home/k896951/Maildir/.nonspam/cur -type f -print0 | xargs -0 -n 1 -P
5 -I{} bogofilter -nvI '{}'
```

/home/k896951/Maildir/.spam はうちの環境でSPAMを集めたメールボックスのフォルダ。実際の開封済みメールは/home/k896951/Maildir/.spam/cur に入っている。
/home/k896951/Maildir/.nonspam は今回のためにNonSPAMなメールをコピーしたメールボックスのフォルダ。実際の開封済みメールは/home/k896951/Maildir/.nonspam/cur に入っている。

bogofilter-sqlite のオプションは、

- -s : これから読むメールはSPAMである
- -n : これから読むメールはNonSPAMである
- -v : 処理中に表示を行う
- -I : 指定の(メール)ファイルを読み込む。

findコマンドとxargsコマンドでbogofilterを5多重起動している。
実際には□SPAM学習とNonSPAM学習を異なるコンソールから実行したので、10多重で実行したことになる。

実行中、

```
# 206 words, 1 message
retrying registration after avoided deadlock...
retrying registration after avoided deadlock...
# 186 words, 1 message
retrying registration after avoided deadlock...
# 232 words, 1 message
# 204 words, 1 message
# 202 words, 1 message
retrying registration after avoided deadlock...
□
□
```

とデッドロック云々のメッセージが出る。これは、データベースファイルwordlist.dbへのアクセスで排

他制御が発生して待ちに入るため。10多重なので、１つのデータベースファイルに10個のプロセスからアクセスが発生するからこれは仕方がない。多重実行しないでシーケンシャルに1メールずつの方が速かったかもしれない。

SPAM/NonSPAMの再学習

ある程度学習が進んだところで、実際にメールを受信し判定させてみる。

うちの環境は、受信したメールを maildrop で仕分けさせている。 以下は maildrop が読み込むルール定義。

```
$ cat ~/.mailfilter
### 配送先などのログを取る場合:
logfile '/home/k896951/maildrop.log'

### spam 判定
xfilter "bogofilter -u -e -p"

### Spamだ！
if (/^X-Bogosity: Spam, tests=bogofilter/:h)
{
  to "/home/k896951/Maildir/.spam/"
}

### 判断に苦しむメール
if (/^X-Bogosity: Unsure, tests=bogofilter/:h)
{
  to "/home/k896951/Maildir/.unsure/"
}

### 大丈夫そうなメール
to "/home/k896951/Maildir/"

$
```

bogofilter-sqliteを通したメールにはメールヘッダ X-Bogosity が付与される。

- X-Bogosity: Ham, ... ← NonSPAMと判定した。メインフォルダに配送する。
- X-Bogosity: Spam, ... ← SPAMと判定したspamフォルダに配送する。
- X-Bogosity: Unsure, ... ← 判定できなかったunsureフォルダに配送する。

このメールヘッダを見てmaildropはメールをそれぞれ指定のメールボックスフォルダに配送する。

メールクライアントで受信したメールを確認すると、いくつかのメールはまだ判定が甘くSPAM判定されずunsureフォルダに溜まる。

Unsure から SPAMと再学習

メールボックスフォルダ unsure から、ユーザから見てSPAMと判断したメールをメールボックスフォルダ NG へ移動しておく。 既に開封されたメールは/home/k896951/Maildir/.NG/cur未開封メール

は/home/k896951/Maildir/.NG/new に格納されている。 事前の学習とオプションは一緒。

```
$ find /home/k896951/Maildir/.NG/cur -type f -print0 | xargs -0 -n 1 -P 5 -
I{} bogofilter -sI '{}'
$ find /home/k896951/Maildir/.NG/new -type f -print0 | xargs -0 -n 1 -P 5 -
I{} bogofilter -sI '{}'
```

Unsure から NonSPAMと再学習

メールボックスフォルダ unsureから、ユーザから見てNonSPAMと判断したメールをメールボックスフォルダ OK へ移動しておく。 既に開封されたメールは/home/k896951/Maildir/.OK/cur□未開封メールは/home/k896951/Maildir/.OK/new に格納されている。 事前の学習とオプションは一緒。

```
$ find /home/k896951/Maildir/.OK/cur -type f -print0 | xargs -0 -n 1 -P 5 -
I{} bogofilter -nI '{}'
$ find /home/k896951/Maildir/.OK/new -type f -print0 | xargs -0 -n 1 -P 5 -
I{} bogofilter -nI '{}'
```

SPAM判定を NonSPAMと再学習

メールボックスフォルダ spam から、ユーザから見てNonSPAMと判断したメールをメールボックスフォルダ NGtoOK へ移動しておく。 既に開封されたメールは/home/k896951/Maildir/.NGtoOK/cur□未開封メールは/home/k896951/Maildir/.NGtoOK/new に格納されている。

bogofilter-sqlite のオプションは、

- -S : SPAMではなかったのでこのメールで登録した情報を無効にする。
- -n : これから読むメールはNonSPAMである
- -I : 指定の(メール)ファイルを読み込む。

つまり□SPAM登録を解除してNonSPAMとして再登録する意味となる。

```
$ find /home/k896951/Maildir/.OK/cur -type f -print0 | xargs -0 -n 1 -P 5 -
I{} bogofilter -SnI '{}'
$ find /home/k896951/Maildir/.OK/new -type f -print0 | xargs -0 -n 1 -P 5 -
I{} bogofilter -SnI '{}'
```

NonSPAM判定を SPAMと再学習

メインメールボックスフォルダから、ユーザから見てSPAMと判断したメールをメールボックスフォルダ OKtoNG へ移動しておく。 既に開封されたメールは/home/k896951/Maildir/.OKtoNG/cur□未開封メールは/home/k896951/Maildir/.OKtoNG/new に格納されている。

bogofilter-sqlite のオプションは、

- -N : NonSPAMではなかったのでこのメールで登録した情報を無効にする。
- -s : これから読むメールはSPAMである
- -I : 指定の(メール)ファイルを読み込む。

つまり□NonSPAM登録を解除してSPAMとして再登録する意味となる。

```
$ find /home/k896951/Maildir/.OK/cur -type f -print0 | xargs -0 -n 1 -P 5 -
I{} bogofilter -NsI '{}'
$ find /home/k896951/Maildir/.OK/new -type f -print0 | xargs -0 -n 1 -P 5 -
I{} bogofilter -NsI '{}'
```

再学習後のメール後処理

再学習後、メールボックスフォルダ OK,NG,OKtoNG,NGtoOK, に残ったメールは、妥当なフォルダへ手作業で戻してもいいし、改めて maildir で配送し直しさせてもしい。

うちではダサいけどこんなシェルスクリプトをcronで定期実行させてる。

```
#!/bin/sh
. /home/k896951/.profile >/dev/null
basedir="/home/k896951/Maildir"

rescan() {
  case $2 in
    "-n"  ) find $basedir/$1 -type f -print0 | xargs -0 -n 1 -P 5 -I{}
bogofilter -nI  '{}' ;;
    "-s"  ) find $basedir/$1 -type f -print0 | xargs -0 -n 1 -P 5 -I{}
bogofilter -sI  '{}' ;;
    "-Sn" ) find $basedir/$1 -type f -print0 | xargs -0 -n 1 -P 5 -I{}
bogofilter -SnI '{}' ;;
    "-Ns" ) find $basedir/$1 -type f -print0 | xargs -0 -n 1 -P 5 -I{}
bogofilter -NsI '{}' ;;
  esac
  sleep 1
  find $basedir/$1 -type f -print0 | xargs -0 -n 1 -I{} sh -c "cat '{}' |
maildrop"
  find $basedir/$1 -type f -print0 | xargs -0      -I{} rm '{}'
}

rescan ".OK/cur"     "-n"
rescan ".OK/new"     "-n"
rescan ".NG/cur"     "-s"
rescan ".NG/new"     "-s"
rescan ".NGtoOK/cur" "-Sn"
rescan ".NGtoOK/new" "-Sn"
rescan ".OKtoNG/cur" "-Ns"
rescan ".OKtoNG/new" "-Ns"
```

このスクリプトは、実行中にメールボックスフォルダのメール数の増減があっても気にしていない。 例えばmaildropの処理実行中に追加でメールが移動されてきても、一切処理されずrmコマンドで消されてしまう。 重要なメールは移動ではなくコピーで処理してオリジナルは残すようにした方がいい。

追記：現在のwordlist.db

2017/02/18 12:00 開始で 2017/02/18 22:45 にNonSPAM学習終了。2017/02/19 12:00 にSPAM学習が完了。

データベースファイルは約270MBのサイズになった。

作成されたデータベースを覗いてみる。

```
$ sqlite3 wordlist.db
SQLite version 3.15.1 2016-11-04 12:08:49
Enter ".help" for usage hints.
sqlite> .databases
seq  name             file
---  ---------------  ----------------------------------------------------------
----
0    main             /home/k896951/.bogofilter/wordlist.db
sqlite> .tables
bogofilter
sqlite> .fullschema
CREATE TABLE bogofilter (   key   BLOB PRIMARY KEY,   value BLOB);
CREATE INDEX bfidx ON bogofilter(key,value);
/* No STAT tables available */
sqlite> .show
        echo: off
         eqp: off
     explain: auto
     headers: off
        mode: list
   nullvalue: ""
      output: stdout
colseparator: "|"
rowseparator: "\n"
       stats: off
       width:
    filename: wordlist.db
sqlite> select * from bogofilter limit 2000000;
.ENDIAN32|^D^C^B^A^D^C^B^A
.WORDLIST_VERSION|4?^A
.ENCODING|^B
rcvd:95.153.177.143|^A
rcvd:x153x177x143.kubangsm.ru|^A
from:psxcfty.com|^A

--- 中略 ---

rcvd:216.215.90.56|^A^M
rcvd:27.36.195.137|^A^M
rtrn:reuyguwxgeefe|^A^M
from:d0205.adgjl.info|^B^M
from:info5324|^B^M
```

```
from:イマが狙い目|^B^M
head:d0205.adgjl.info|^B^M
head:info5324|^B^M
rcvd:180.131.26.206|^A^M
rcvd:d0205.adgjl.info|^B^M
rtrn:d0205.adgjl.info|^B^M
rtrn:info5324|^B^M
from:hczcmt|^A^M
from:molnlycke.net|^A^M
rcvd:119.48.196.23|^A^M
rtrn:hczcmt|^A^M
rtrn:molnlycke.net|^A^M
rcvd:111.181.208.170|^A^M
rcvd:250.200.76.188|^A^M
VEGAS□|^B^M
from:info623|^B^M
from:j0235.zcbmv.info|^B^M
head:info623|^B^M
head:j0235.zcbmv.info|^B^M
rcvd:180.131.26.236|^A^M
rcvd:j0235.zcbmv.info|^B^M
rtrn:info623|^B^M
rtrn:j0235.zcbmv.info|^B^M
subj:今が狙い目！□|^B^M
□□□□□318|^B^M
□□□□□545|^B^M
□□□□□615|^B^M
□□□□□75|^B^M
【アップル|^B^M
【ヤスオク|^B^M
ホワイト□|^B^M
from:approval79k|^A^M
rcvd:76.108.207.139|^A^M
rtrn:thurman4qd|^A^M
from:hlrjt|^A^M
head:hlrjt|^A^M
rcvd:144.110.175.191|^A^M
rcvd:219.138.218.39|^A^M
rtrn:hlrjt|^A^M
Walker□□□□□□□□|^A^M
from:Girls|^A^M
from:girls.walker|^A^M
head:girls.walker|^A^M
rcvd:112.196.254.82|^A^M
rtrn:girls.walker|^A^M
subj:□重要□Girls|^A^M
□□□□□□Girls|^A^M
from:あやか|^A^M
head:ayaka|^A^M
rcvd:180.232.107.88|^A^M
rtrn:ayaka|^A^M
```

```
from:R.x|^A^M
from:elap.it|^A^M
from:trinfo|^A^M
rcvd:162.178.56.94|^A^M
rcvd:41.136.201.186|^A^M
rtrn:elap.it|^A^M
rtrn:trinfo|^A^M
www.pillswrg.com|^B^M
from:xabkwd.com|^E^M
head:xabkwd.com|^E^M
rcvd:180.232.107.250|^A^M
rcvd:xabkwd.com|^E^M
rtrn:xabkwd.com|^E^M
xabkwd.com|^E^M
rcvd:14.105.214.86|^A^M
rcvd:96.128.16.44|^A^M
head:errv51555_1061|^A^M
rtrn:errv51555_1061|^A^M

--- 以下略 ---
```

テーブルとインデクスが1つずつ、テーブルはキーもデータもBLOB型。 valueは読めないけどkeyは読める。

keyについては " 出現ヶ所:キーワード " の組み合わせにみえる。 メールヘッダや本文情報毎にKVSのような使い方をしているのかな、と推測Valueの方が不明なので何とも言えないけど。

ついでなので、最近頓に多いISPの名前で検索をしてみた。

```
sqlite> select "key" from bogofilter where "key" like "%:%biglobe%" order by
1;
from:bcs.biglobe.ne.jp
from:bma.biglobe.ne.jp
from:kdn.biglobe.ne.jp
from:kfa.biglobe.ne.jp
from:kjc.biglobe.ne.jp
from:knh.biglobe.ne.jp
from:kni.biglobe.ne.jp
from:kpb.biglobe.ne.jp
from:ksf.biglobe.ne.jp
from:kub.biglobe.ne.jp
from:kyf.biglobe.ne.jp
from:kza.biglobe.ne.jp
from:mpd.biglobe.ne.jp
from:mrg.biglobe.ne.jp
from:msc.biglobe.ne.jp
from:msd.biglobe.ne.jp
from:mth.biglobe.ne.jp
from:muc.biglobe.ne.jp
from:mvi.biglobe.ne.jp
from:mwb.biglobe.ne.jp
```

```
from:ok.biglobe.ne.jp
from:one.biglobe.ne.jp
from:oyaji.biglobe.ne.jp
from:pan.biglobe.ne.jp
from:peace.biglobe.ne.jp
from:pearl.biglobe.ne.jp
from:pegasus.biglobe.ne.jp
from:piyo.biglobe.ne.jp
from:pmb.biglobe.ne.jp
from:pretty.biglobe.ne.jp
from:pride.biglobe.ne.jp
from:private.biglobe.ne.jp
from:puppy.biglobe.ne.jp
from:red.biglobe.ne.jp
from:river.biglobe.ne.jp
from:robin.biglobe.ne.jp
from:rose.biglobe.ne.jp
from:ruby.biglobe.ne.jp
from:s.biglobe.ne.jp
from:sakura.biglobe.ne.jp
from:sapphire.biglobe.ne.jp
from:sea.biglobe.ne.jp
from:seven.biglobe.ne.jp
from:sheep.biglobe.ne.jp
from:silver.biglobe.ne.jp
from:sky.biglobe.ne.jp
from:snowy.biglobe.ne.jp
from:soccer.biglobe.ne.jp
from:spice.biglobe.ne.jp
from:star.biglobe.ne.jp
from:sunshine.biglobe.ne.jp
from:surfing.biglobe.ne.jp
from:sweet.biglobe.ne.jp
from:swing.biglobe.ne.jp
from:symphony.biglobe.ne.jp
from:tea.biglobe.ne.jp
from:techno.biglobe.ne.jp
from:tennis.biglobe.ne.jp
from:thanks.biglobe.ne.jp
from:tiny.biglobe.ne.jp
from:topaz.biglobe.ne.jp
from:tulip.biglobe.ne.jp
from:ultra.biglobe.ne.jp
from:valuestar.biglobe.ne.jp
from:venus.biglobe.ne.jp
from:violet.biglobe.ne.jp
from:viva.biglobe.ne.jp
from:wonder.biglobe.ne.jp
from:xqd.biglobe.ne.jp
from:xqe.biglobe.ne.jp
from:xqg.biglobe.ne.jp
```

```
from:xqj.biglobe.ne.jp
from:xug.biglobe.ne.jp
from:xvg.biglobe.ne.jp
from:y.biglobe.ne.jp
from:yours.biglobe.ne.jp
from:z.biglobe.ne.jp
from:zoo.biglobe.ne.jp
from:zzz.biglobe.ne.jp
head:X-Biglobe-Sender
head:X-Biglobe-Spnum
head:X-Biglobe-VirusCheck
head:bcs.biglobe.ne.jp
head:bma.biglobe.ne.jp
head:kfa.biglobe.ne.jp
head:kki.biglobe.ne.jp
head:kni.biglobe.ne.jp
head:kpb.biglobe.ne.jp
head:kss.biglobe.ne.jp
head:kub.biglobe.ne.jp
head:kyf.biglobe.ne.jp
head:kza.biglobe.ne.jp
head:mpd.biglobe.ne.jp
head:mrg.biglobe.ne.jp
head:msc.biglobe.ne.jp
head:msd.biglobe.ne.jp
head:muc.biglobe.ne.jp
head:mvi.biglobe.ne.jp
head:mwb.biglobe.ne.jp
head:rcpt-expgw.biglobe.ne.jp
head:replybiglobe
head:spfkv6.biglobe.ne.jp
head:valuestar.biglobe.ne.jp
rcvd:biglobe.click
rcvd:bpkv.mvi.biglobe.ne.jp
rcvd:kfa.biglobe.ne.jp
rcvd:kza.biglobe.ne.jp
rcvd:mrg.biglobe.ne.jp
rcvd:msd.biglobe.ne.jp
rcvd:rcpt-expgw.biglobe.ne.jp
rcvd:replybiglobe
rcvd:smtp-gw.biglobe.ne.jp
rcvd:vc-gw.biglobe.ne.jp
rtrn:bcs.biglobe.ne.jp
rtrn:biglobe.click
rtrn:bma.biglobe.ne.jp
rtrn:kfa.biglobe.ne.jp
rtrn:kki.biglobe.ne.jp
rtrn:kni.biglobe.ne.jp
rtrn:kss.biglobe.ne.jp
rtrn:kub.biglobe.ne.jp
rtrn:kyf.biglobe.ne.jp
```

```
rtrn:kza.biglobe.ne.jp
rtrn:mpd.biglobe.ne.jp
rtrn:mrg.biglobe.ne.jp
rtrn:msc.biglobe.ne.jp
rtrn:msd.biglobe.ne.jp
rtrn:muc.biglobe.ne.jp
rtrn:mvi.biglobe.ne.jp
rtrn:mwb.biglobe.ne.jp
rtrn:ok.biglobe.ne.jp
rtrn:one.biglobe.ne.jp
rtrn:oyaji.biglobe.ne.jp
rtrn:pan.biglobe.ne.jp
rtrn:peace.biglobe.ne.jp
rtrn:pearl.biglobe.ne.jp
rtrn:pegasus.biglobe.ne.jp
rtrn:piyo.biglobe.ne.jp
rtrn:pmb.biglobe.ne.jp
rtrn:pretty.biglobe.ne.jp
rtrn:pride.biglobe.ne.jp
rtrn:private.biglobe.ne.jp
rtrn:puppy.biglobe.ne.jp
rtrn:red.biglobe.ne.jp
rtrn:river.biglobe.ne.jp
rtrn:robin.biglobe.ne.jp
rtrn:rose.biglobe.ne.jp
rtrn:ruby.biglobe.ne.jp
rtrn:s.biglobe.ne.jp
rtrn:sakura.biglobe.ne.jp
rtrn:sapphire.biglobe.ne.jp
rtrn:sea.biglobe.ne.jp
rtrn:seven.biglobe.ne.jp
rtrn:sheep.biglobe.ne.jp
rtrn:silver.biglobe.ne.jp
rtrn:sky.biglobe.ne.jp
rtrn:snowy.biglobe.ne.jp
rtrn:soccer.biglobe.ne.jp
rtrn:spice.biglobe.ne.jp
rtrn:star.biglobe.ne.jp
rtrn:sunshine.biglobe.ne.jp
rtrn:surfing.biglobe.ne.jp
rtrn:sweet.biglobe.ne.jp
rtrn:swing.biglobe.ne.jp
rtrn:symphony.biglobe.ne.jp
rtrn:tea.biglobe.ne.jp
rtrn:techno.biglobe.ne.jp
rtrn:tennis.biglobe.ne.jp
rtrn:thanks.biglobe.ne.jp
rtrn:tiny.biglobe.ne.jp
rtrn:topaz.biglobe.ne.jp
rtrn:tulip.biglobe.ne.jp
rtrn:ultra.biglobe.ne.jp
```

```
rtrn:valuestar.biglobe.ne.jp
rtrn:venus.biglobe.ne.jp
rtrn:violet.biglobe.ne.jp
rtrn:viva.biglobe.ne.jp
rtrn:wonder.biglobe.ne.jp
rtrn:xqd.biglobe.ne.jp
rtrn:xqe.biglobe.ne.jp
rtrn:xqg.biglobe.ne.jp
rtrn:xqj.biglobe.ne.jp
rtrn:xug.biglobe.ne.jp
rtrn:xvg.biglobe.ne.jp
rtrn:y.biglobe.ne.jp
rtrn:yours.biglobe.ne.jp
rtrn:z.biglobe.ne.jp
rtrn:zoo.biglobe.ne.jp
rtrn:zzz.biglobe.ne.jp
to:msd.biglobe.ne.jp
to:replybiglobe
sqlite>
```

結構**サブドメインあるみたい**。

技術資料, bogofilter, bogofilter-sqlite, maildrop, mail