By Falko Timme

Published: 2007-02-12 17:45

Fight Image Spam With FuzzyOCR And SpamAssassin On Debian/Ubuntu

Version 1.0
Author: Falko Timme <ft [at] falkotimme [dot] com>
Last edited 02/12/2007

This tutorial describes how to scan emails for image spam with **FuzzyOCR**. FuzzyOCR is a plugin for SpamAssassin which is aimed at unsolicited bulk mail containing images as the main content carrier. Using different methods, it analyzes the content and properties of images to distinguish between normal mails (ham) and spam mails. FuzzyOCR tries to keep the system load low by scanning only mails that have not already been categorized as spam by SpamAssassin, thus avoiding unnecessary work.

I do not issue any guarantee that this will work for you!

1 Preliminary Note

In this article I will use Debian Etch for the base system. The steps to install FuzzyOCR should be the same for Ubuntu systems.

I assume that SpamAssassin is already installed and working, with *`/etc/mail/spamassassin/`* as its main configuration directory. If your directory is different (e.g. if you have **ISPConfig** installed, the directory is *`/home/admispconfig/ispconfig/tools/spamassassin/etc/mail/spamassassin/`*), this is no problem. I will annotate where to change what.

Please make sure that your SpamAssassin version works with FuzzyOCR. For example, the FuzzyOCR version I'm going to install here (*`fuzzyocr-3.5.1-devel.tar.gz`*) requires SpamAssassin 3.1.4 or newer.

2 Install The Prerequisites For FuzzyOCR

FuzzyOCR has some prerequisites like *`ocrad`* and *`gocr`* that we can install like this:

```
apt-get install netpbm gifsicle libungif-bin gocr ocrad libstring-approx-perl libmldbm-sync-perl imagemagick tesseract-ocr
```

3 Install FuzzyOCR

Next we download and install the latest FuzzyOCR devel version from http://fuzzyocr.own-hero.net/wiki/Downloads. We download the devel version instead of the stable version because the FuzzyOCR developers say:

"The current recommendation is the development version because the stable version lacks features and is very old."

```
cd /usr/src/

wget http://users.own-hero.net/~decoder/fuzzyocr/fuzzyocr-3.5.1-devel.tar.gz
```

Then we unpack FuzzyOCR and move all `FuzzyOcr*` files and the `FuzzyOcr` directory (they are all in the `FuzzyOcr-3.5.1/` directory) to `/etc/mail/spamassassin`:

```
tar xvfz fuzzyocr-3.5.1-devel.tar.gz

cd FuzzyOcr-3.5.1/

mv FuzzyOcr* /etc/mail/spamassassin/
```

If your SpamAssassin directory is different, e.g. `/home/admispconfig/ispconfig/tools/spamassassin/etc/mail/spamassassin/`, then the last command should be replaced with

```
mv FuzzyOcr* /home/admispconfig/ispconfig/tools/spamassassin/etc/mail/spamassassin/
```

Don't delete the `/usr/src/FuzzyOcr-3.5.1/` directory yet, there's a directory with sample image spam emails in there (`samples/`) that we need later on to test if FuzzyOCR is working as expected.

So FuzzyOCR is now installed, now we need to configure it.

4 Configure FuzzyOCR

FuzzyOCR's configuration file is `/etc/mail/spamassassin/FuzzyOcr.cf`. In that file almost everything is commented out. We open that file now and make some modifications:

```
vi /etc/mail/spamassassin/FuzzyOcr.cf
```

Put the following line into it to define the location of FuzzyOCR's spam words file:

```
[...]
focr_global_wordlist /etc/mail/spamassassin/FuzzyOcr.words
[...]
```

`/etc/mail/spamassassin/FuzzyOcr.words` is a predefined word list that comes with FuzzyOCR. You can adjust it to your needs if you like.

Next change

```
[...]
# Include additional scanner/preprocessor commands here:
#
focr_bin_helper pnmnorm, pnminvert, pamthreshold, ppmtopgm, pamtopnm
focr_bin_helper tesseract
[...]
```

to

```
[...]
# Include additional scanner/preprocessor commands here:
#
focr_bin_helper pnmnorm, pnminvert, convert, ppmtopgm, tesseract
[...]
```

Finally add/enable the following lines:

```
[...]
# Search path for locating helper applications
focr_path_bin /usr/local/netpbm/bin:/usr/local/bin:/usr/bin

focr_preprocessor_file /etc/mail/spamassassin/FuzzyOcr.preps
focr_scanset_file /etc/mail/spamassassin/FuzzyOcr.scansets

focr_enable_image_hashing 2
focr_digest_db /etc/mail/spamassassin/FuzzyOcr.hashdb
focr_db_hash /etc/mail/spamassassin/FuzzyOcr.db
focr_db_safe /etc/mail/spamassassin/FuzzyOcr.safe.db
[...]
```

With the last four lines you enable image hashing. This is what the FuzzyOCR developers say about image hashing:

"The Image hashing database feature allows the plugin to store a vector of image features to a database, so it knows this image when it arrives a second time (and therefore does not need to scan it again). The special thing about this function is that it also recognizes the image again if it was changed slightly (which is done by spammers). "

If you use `/home/admispconfig/ispconfig/tools/spamassassin/etc/mail/spamassassin` instead of `/etc/mail/spamassassin`, FuzzyOCR's configuration file is `/home/admispconfig/ispconfig/tools/spamassassin/etc/mail/spamassassin/FuzzyOcr.cf` instead of `/etc/mail/spamassassin/FuzzyOcr.cf`, so edit that one. In the configuration file you can now either replace all occurrences of

`/etc/mail/spamassassin` with `/home/admispconfig/ispconfig/tools/spamassassin/etc/mail/spamassassin`, you leave it as shown before and create a symlink from `/etc/mail/spamassassin` to `/home/admispconfig/ispconfig/tools/spamassassin/etc/mail/spamassassin` like this:

```
mkdir /etc/mail/

ln -s /home/admispconfig/ispconfig/tools/spamassassin/etc/mail/spamassassin/ /etc/mail/spamassassin
```

That's it already for the FuzzyOCR configuration. Now let's see if it works as expected.

5 Test FuzzyOCR

I mentioned before that FuzzyOCR comes with sample image spam mails (in the `samples/` directory):

```
ls -l /usr/src/FuzzyOcr-3.5.1/samples/
```

The output should look like this:

```
total 156
  -rw-r--r-- 1 administrator users 13633 2007-01-07 12:55 ocr-animated.eml
  -rw-r--r-- 1 administrator users 16108 2007-01-07 12:55 ocr-gif.eml
  -rw-r--r-- 1 administrator users 27506 2007-01-07 12:55 ocr-jpg.eml
  -rw-r--r-- 1 administrator users 27842 2007-01-07 12:59 ocr-multi.eml
  -rw-r--r-- 1 administrator users 24657 2007-01-07 12:55 ocr-obfuscated.eml
  -rw-r--r-- 1 administrator users 18236 2007-01-07 12:55 ocr-png.eml
  -rw-r--r-- 1 administrator users 16113 2007-01-07 12:55 ocr-wrongext.eml
-rw-r--r-- 1 administrator users 3576 2007-01-07 12:55 README
```

We can feed each of these emails to SpamAssassin now to see if FuzzyOCR is linked correctly into SpamAssassin. Find out where your `spamassassin` executable is (normally it's in your `PATH` - you can find out if this is the case by running

```
which spamassassin
```

If it shows a result, `spamassassin` is in your `PATH`, and you don't need to specify the full path to `spamassassin` to run it.)

If you don't know where `spamassassin` is, you can find out by running

```
updatedb

locate spamassassin
```

If you use ISPConfig, `spamassassin` is here: `/home/admispconfig/ispconfig/tools/spamassassin/usr/bin/spamassassin`

Now that you know where `spamassassin` is, you can feed the sample image spam mails to `spamassassin` like this:

```
/path/to/spamassassin --debug FuzzyOcr < /usr/src/FuzzyOcr-3.5.1/samples/ocr-gif.eml > /dev/null
```

E.g.

```
/home/admispconfig/ispconfig/tools/spamassassin/usr/bin/spamassassin --debug FuzzyOcr < /usr/src/FuzzyOcr-3.5.1/samples/ocr-gif.eml > /dev/null
```

or, if `spamassassin` is in your `PATH`:

```
spamassassin --debug FuzzyOcr < /usr/src/FuzzyOcr-3.5.1/samples/ocr-gif.eml > /dev/null
```

You should now see a lot of output, the end should look like this:

```
[...]
  [10025] dbg: FuzzyOcr:
  [10025] dbg: FuzzyOcr: Friday Augurt 4, 4:01 pm ET
  [10025] dbg: FuzzyOcr: LAS VEGAS, NEVADA--(MARKET WIRE)--Aug 4, 2006 -- auantum Energy, lnc. (OTC
  [10025] dbg: FuzzyOcr: BB:aEGY.oB-_-
  [10025] dbg: FuzzyOcr: auantum Energy, lnc. is pleased to announce that it has applied to have its shares listed for
  [10025] dbg: FuzzyOcr: trading on the Frankfurt Stock Exchange. The company has retained the services ofBaltic
  [10025] dbg: FuzzyOcr: lnvestment Group of Hamburg, Germany to assist with the application.
  [10025] dbg: FuzzyOcr:
  [10025] dbg: FuzzyOcr: _ qEGY,OB "
  [10025] dbg: FuzzyOcr:
  [10025] dbg: FuzzyOcr: <<=end
  [10025] info: FuzzyOcr: Scanset "ocrad" found word "target" with fuzz of 0.0000
  [10025] info: FuzzyOcr: line: "short term price target oo"
  [10025] info: FuzzyOcr: Scanset "ocrad" found word "service" with fuzz of 0.0000
  [10025] info: FuzzyOcr: line: "trading on the frankfurt stock exchange the company has retained the services ofbaltic"
  [10025] info: FuzzyOcr: Scanset "ocrad" found word "stock" with fuzz of 0.0000
  [10025] info: FuzzyOcr: line: "hot energy stocki"
  [10025] info: FuzzyOcr: Scanset "ocrad" found word "stock" with fuzz of 0.0000
  [10025] info: FuzzyOcr: line: "trading on the frankfurt stock exchange the company has retained the services ofbaltic"
  [10025] info: FuzzyOcr: Scanset "ocrad" found word "price" with fuzz of 0.0000
  [10025] info: FuzzyOcr: line: "current price o"
  [10025] info: FuzzyOcr: Scanset "ocrad" found word "price" with fuzz of 0.0000
  [10025] info: FuzzyOcr: line: "short term price target oo"
  [10025] info: FuzzyOcr: Scanset "ocrad" found word "company" with fuzz of 0.0000
  [10025] info: FuzzyOcr: line: "trading on the frankfurt stock exchange the company has retained the services ofbaltic"
  [10025] info: FuzzyOcr: Scanset "ocrad" found word "recommendation" with fuzz of 0.0000
  [10025] info: FuzzyOcr: line: "sboog bup recommendation"
  [10025] dbg: FuzzyOcr: Enough OCR Hits without space stripping, skipping second matching pass...
  [10025] info: FuzzyOcr: Scanset "ocrad" generates enough hits (8), skipping further scansets...
  [10025] info: FuzzyOcr: Message is spam, score = 15.000
  [10025] info: FuzzyOcr: Adding Hash to "/etc/mail/spamassassin/FuzzyOcr.db" with score "15.000"
  [10025] dbg: FuzzyOcr: Digest:
538584:327:549:7::255:255:255:255:168580::0:0:0:0:9098::0:128:0:75:1086::0:0:128:15:395::128:0:128:53:213::0:0:255:29:115
  [10025] info: FuzzyOcr: Words found:
  [10025] info: FuzzyOcr: "target" in 1 lines
  [10025] info: FuzzyOcr: "service" in 1 lines
  [10025] info: FuzzyOcr: "stock" in 2 lines
  [10025] info: FuzzyOcr: "price" in 2 lines
  [10025] info: FuzzyOcr: "company" in 1 lines
  [10025] info: FuzzyOcr: "recommendation" in 1 lines
  [10025] info: FuzzyOcr: (12 word occurrences found)
  [10025] dbg: FuzzyOcr: Remove DIR: /tmp/.spamassassin10025QnPTq8tmp
  [10025] dbg: FuzzyOcr: FuzzyOcr ending successfully...
[10025] dbg: FuzzyOcr: Processed in 2.191381 sec.
```

As you see `/usr/src/FuzzyOcr-3.5.1/samples/ocr-gif.eml` has been categorized as spam with a score of 15 points, so FuzzyOCR is working.

So your SpamAssassin is now able to recognize image spam thanks to the help of FuzzyOCR.

6 Links

- FuzzyOCR: ***http://www.fuzzyocr.net***
- SpamAssassin: ***http://spamassassin.apache.org***
- Debian: ***http://www.debian.org***
- Ubuntu: ***http://www.ubuntu.com***