FuzzyOcrPlugin

Attention

"This project is UNMAINTAINED as of 2009-06-01. Use it at your own risk."

Informations on this page are out-of-date and need updating. Especially Requirements, Configuration and Installation Instructions might not be accurate anymore for current versions.

FuzzyOcr now has a webpage located at:

http://fuzzyocr.own-hero.net/

More up-to-date informations will be available there soon, installation instructions for current versions are shipped with the tarballs available there.

How it works

NOTE: This plugin is based on the OcrPlugin written by Maarten de Boer and was extended and improved.

This plugin checks for specific keywords in image/gif, image/jpeg or image/png attachments, using `gocr` (an *optical character recognition* program).

This plugin can be used to detect spam that puts all the real spam content in an attached image. The mail itself only random text and random html, without any URL's or identifiable information.

Additionally to the normal OcrPlugin, it can do approximate matches on words, so errors in recognition or attempts to obfuscate the text inside the image will not cause the detection to fail. Another improvement was to move the wordlist into the configuration file so it can be easily extended.

Requirements

You will need `giftopnm, jpegtopnm and pngtopnm` (from netpbm), imagemagick and `gocr` installed.

Additionally, you will need the perl module

```
String::Approx
```

and several tools from

```
giflib
```

(also known as libungif).

ATTENTION: There has been a segfault discovered in both

```
giftext
```

and

```
gocr
```

Patches for the sources are to be found in the download directory of FuzzyOcr. Not using these can make problems under certain circumstances.

Notes for Fedora Core 5 (or higher) users: The package giflib-utils provides giffix. The package netpbm-progs provides giftopnm, etc.

Notes for other Redhat/FC users: The packages libungif and libungif-progs should be installed to provide giffix.

Notes for Debian users: The package libungif-bin provides giffix.

Attention when using RedHat! The gocr RPM for RedHat is faulty and causes bad recognition results. Here is a quote from Ken Bass:

```
I installed gocr 0.40 using an RPM / (source RPM). For whatever reason,
the RPM configures the gocr using 'configure --with-netpbm=no'. This
netpbm=no option causes some images to not be decoded properly. I get
much more garbage. I had to rebuild/reinstall gocr 0.40 without
disabling netpbm for best results.

I've reported this to the gocr mailing list, sent him a sample image,
and hopefully it will be fixed. He seemed to think it would be fixed in
gocr, but I have no timeframe. In the meantime, those with problems,
trying modiying the RPM gocr.spec file or build/install by hand.
```

Changelog

Version 2.0:

- Replaced imagemagick with netpbm tools
- Plugin invokes giffix now on gifs to handle intentionally corrupted gifs
- Added png support
- Added magic byte detection to detect correct file format independantly from content-type
- Added 3 verbosity levels
- Added configuration option for tmp file path and scores

Version 2.1:

- Added scoring for wrong content-type
- Added scoring for broken gif images
- Added configuration for helper applications
- Added autodisable_score feature to disable the OCR engine if the message has already enough points

Version 2.2

- Several bugfixes
- New debug system
- Logfile support
- Proper error handling for most errors

Version 2.3

- Multiple scans with different pnm preprocessing and gocr arguments possible
- Support for interlaced gifs
- Support for animated gifs
- Temporary file handling reorganized
- External wordlist support
- Personalized wordlist support
- Spaces are now stripped from wordlist words and OCR results before matching
- Experimental MD5 Database feature

Installation

Attention: If you need help installing this plugin or have other questions, please use the mailinglist created for this plugin or contact me on IRC (see the end of this page for more informations)

It can be found at http://lists.own-hero.net/mailman/listinfo/devel-spam

Since version 2.3, the tarball contains an INSTALL file and a FAQ file. Both should be read for instructions installing it.

The following informations are a bit older and might not be accurate anymore for version 2.3. Most new parameters are not mentioned here anymore.

Download the tarball (see How to Obtain) to your spamassassin configuration directory and unpack it to /etc/mail/spamassassin/ (You may choose another location but all necessary adjustments to the configuration file are up to you then). Open `FuzzyOcr.cf` and extend the wordlist as you wish. If you have the helper binaries in a different location than the default in the config file specifies, then change these to the correct path.

The scoring is dynamic, more word matches lead to a higher score. The scoring is done as soon as `focr_counts_required` matches were found. It scores exactly `focr_base_score` points then. For every additional match, it scores additionally `focr_add_score` points.

Attention: Do not add a score line to the config file. It will not be used! Scoring is done INTERNALLY and can only be configured with the two parameters described above.

The variable $countreq can be adjusted via the configuration file parameter `focr_counts_required` and indicates the number of matches that need to be found before any score will be triggered.

The variable $threshold is similarly adjusted with the configuration file parameter `focr_threshold`. This is a float value between 0 and 1 and indicates the maximum relative edit distance between the wordlist word and the obfuscated version (less means the words need to be more similar, 0 means identical). The default of 0.3 normally does not need any change. Note that this module also matches substrings (see example).

Explanation of the additional options:

`focr_logfile` - String determining the file to send log messages to. Make sure this is writable!

`focr_verbose` - Verbose level (0 - 2). (1 is currently the default)

- 0 means normal operation.
- 1 means output all words and the corresponding measured distance in the rule output:

```
                        6.0 FUZZY_OCR        BODY: Mail contains an image with common spam text inside
                            Words found:
                            "viagra" with fuzz of 0.2
                            "cialis" with fuzz of 0
                            "viagra" with fuzz of 0.2
                            "levitra" with fuzz of 0
                            (4 word occurrences found)
```

- 2 means same as 1 with an additional output to the logfile (more messages) and temporary files don't get deleted (so you can inspect them)

`focr_bin_*` - Tells the plugin about the helper applications, change to the full path + binary name if your applications are not found.

`focr_wrongctype_score` - Score to give for a wrong content-type (e.g. Image is GIF but content-type says image/jpeg)

`focr_corrupt_score` - Score to give for a corrupted image (Currently only used with GIF images)

`focr_autodisable_score` - If the message has already more points than this value, then the plugin will cancel all further OCR checking.

Example of work

Lets say you have defined `focr_word investor` in your configuration. Now you receive an image which, after converted and recognized gives you:

`ATTENTION ALL IN\lESTORS AND DAY TRADERS`

Then the plugin will find the word investor. It would even succeed if the text was `ATTENTION ALL STUPUDIN\lESTORSHAHA` or `INVSTORSZ` etc.

Generally, the plugin follows these rules:

- The case is not relevant
- All special characters, spaces or numbers are stripped before any matching is done
- Your wordlist word will be found even if it is inside another word (submatching)
- The distance is calculated from the amount of character additions, deletions and substitutions, that need to be done.

Remarks

- The words checked for are specific for some spam I received a lot of recently.
- `gocr` can take up quite a bit of resources, so be careful. But it is only executed for messages that contain gif, png or jpeg attachments.

ToDo

- Rework animated gif handling
- Replace plain MD5 database with a DBM file

– Author: Christian Holler, decoder_at_own-hero_dot_net

How to obtain

You can download the latest tarball containing the `FuzzyOcr.pm` and `FuzzyOcr.cf` from http://users.own-hero.net/~decoder/fuzzyocr/

For support, you can write me an email, or catch me on IRC: Server: irc.own-hero.net Channel: #nmg (My nick is decoder😛)

You can also subscribe to the mailing list found at http://lists.own-hero.net/mailman/listinfo/devel-spam but please understand that it is private since we also talk about development there. If you want to test the newest alpha releases, this is the place were you'd want to be 😉