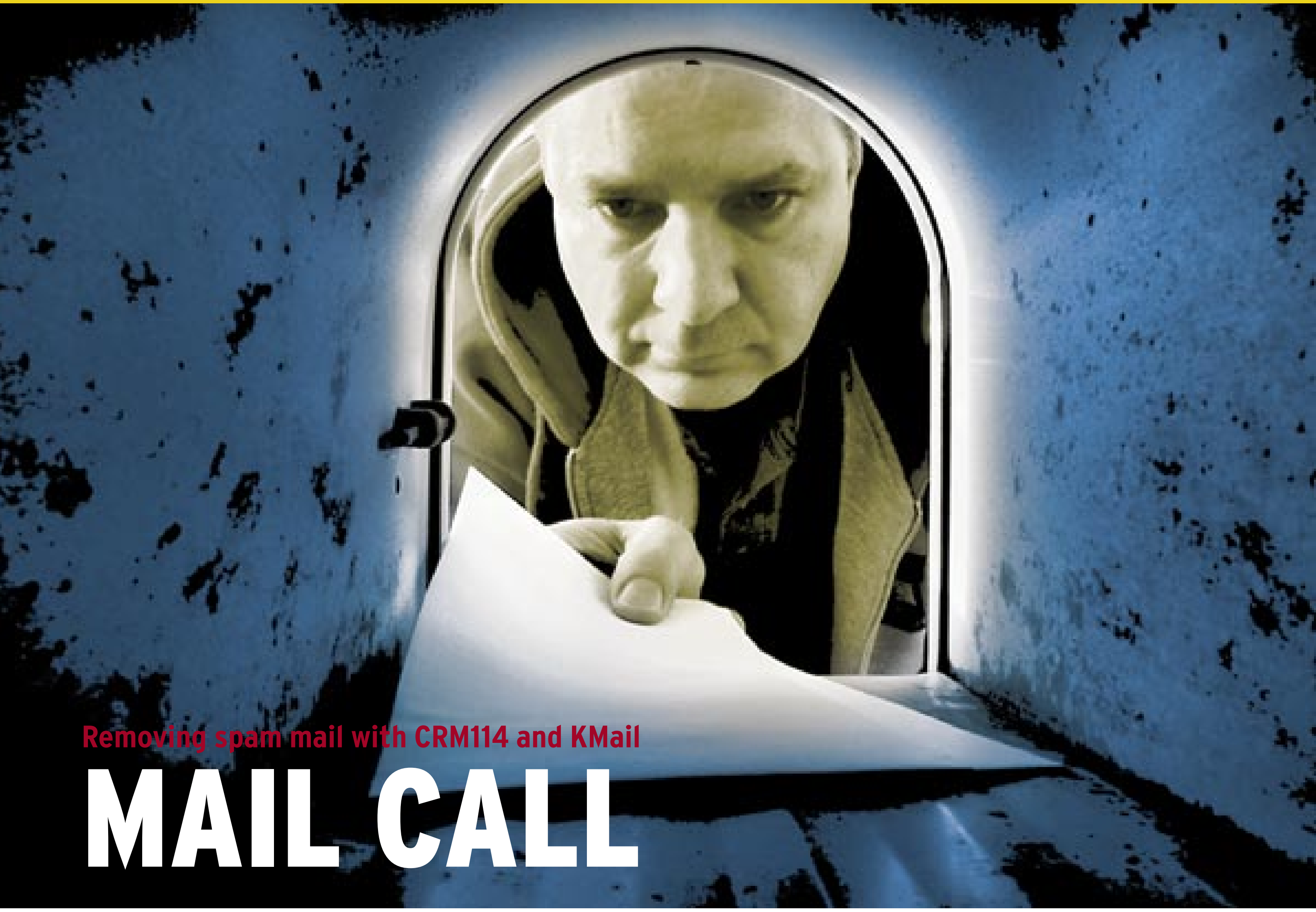


Removing spam mail with CRM114 and KMail

MAIL CALL

The CRM114 filter program, which is integrated with KMail, helps fight spam in POP3 and IMAP mailboxes. Because this flexible filter does not require server-side changes, it is a good choice for users without root access to their mail servers. **BY MARTIN STEIGERWALD**

As if the flood of normal spam isn't bad enough, spammers are now aggravating large parts of the literate world with image-based spam and with messages that don't contain any meaningful content at all. This new spam is a challenge for even the most sophisticated of filters. For example, the database in our lab using the Bogofilter program grew to a size of 100 MB. Receiving mail via KMail on an IBM ThinkPad T23, with a Pentium 3 1.13 GHz CPU and 768 MB RAM, became a time-consuming experience. The KMail interface freezes for the most part, as it does not call the spam filter program as a background task as of this writing [1].

It seems unlikely that the developers will find a solution to this issue before the KDE 4 release, since it would mean rewriting major parts of the program. At the same time, much junk mail remains unfiltered due to image spam and messages with meaningless content. One thing is for sure, I needed a more powerful solution.

CRM 114 Discriminator

For the current lab, I opted for CRM114, an email-sorting program praised by many members of the community [2]. The acronym expands to Controllable Regex Mutilator and refers to a piece of radio reception equipment (the CRM114 Discriminator from Stanley Kubrick's movie "Dr. Strangelove"), which prevents reception of non-authenticated transmissions [3]. Consider it to be a controllable regex modifier if you prefer something less hostile sounding.

CRM114 works like a programming language designed to filter data based on regular expressions. The data can be emails, logfiles or other sources. CRM114 can learn filtering criteria.

CRM114 filters email and has a number of advantages – it is fast and does not need much in the way of resources. It learns quickly and is extremely accurate after initial teaching. Its ambitious goal is to filter email so effectively that spamming no longer makes economic sense for spammers.

Like many other spam filters, CRM114 is designed to be used flexibly. For example, there is nothing to stop you from setting CRM114 up on a mail server with Procmail, Maildrop or using a *.forward* file or integrating the program with the Squirrelmail webmailer or the text-based Mutt client [4][5].

In this article, we will be looking at the basic steps for installing CRM114 and integrating the program with KMail. This method is useful for **POP3** and **IMAP** mailboxes and does not require any server-side changes, which makes this a good choice for users who do not have root access to their own mail servers. If you use IMAP on your own server, a server-side installation would make more sense.

Installing CRM114

The first step is to install and configure CRM114 and the CRM114 mail filter program [4]. Users with Debian or Ubuntu can simply install the *crm114* package. In our lab, we used the *20060704a* version that comes with Debian Etch, which is also included with Ubuntu Edgy and available as a backport for Sarge [6].

If you have Suse or prefer to use a more recent version of CRM114, you will need the **upstream version**. The easiest approach is to use the **statically** linked binary version [7]. Unpack the archive by entering tar -xvzf archivefile, change to the directory created by this step, enter the following command:

```
su -c "make install_binary_only"
```

and provide your root password. You can type *su -c "make uninstall* to uninstall CRM114 whenever needed. For instructions on installing the software from the source code, refer to the Building CRM114 box.

As an initial test, enter *crm -v* to output the CRM114 version number. The following command launches the famous “Hello World” program:

```
crm '-{ output /Hello, world!\n/}'
```

CRM114 will typically wait for an input stream before running a program that creates the corresponding output. Press [Ctrl] + [D] to terminate the input stream and run the program.

Setting Up the Mail Filter

The CRM114 mail filter program parses a configuration file, multiple lists (such as a blacklist and a whitelist), and a **hash** with a default size of 12MB that defines spam and **ham**.

Give the *mkdir ~/.crm114* command to create a directory for the CRM114 files below your home directory, and then type *cd ~/.crm114* to change to that directory. Users with Debian or Ubuntu can then type:

```
cp -a /usr/share/crm114/*.crm .
```

to copy the program files. You will need to copy *mailreaver.crm*, *maillib.crm*, *mailtrainer.crm*, and *rewriteutil.crm* – as a minimum – to the new directory.

If you opted for the upstream version, you can use the files from the binary or source code archive. The current filter program for CRM114 is called *mailreaver.crm*. Do not use the older *mailfilter.crm* program.

You also need to copy the mail filter configuration file, *mailfilter.cf*, to your CRM114 directory and use a text editor to customize the file. For KMail integration, you do not need remote access to CRM114. You can leave this line:

```
:spw: /DEFAULT_PASSWORD/
```

unchanged, despite the warnings.

Select the correct decoder for MIME-encoded mail attachments by removing the pound sign at the start of the *:mime_decoder:* line and commenting any unwanted *:mime_decoder:* lines as needed. Use the following command to find out which MIME encoder is installed on your system:

```
which /usr/bin/mimencode
```

Replace */usr/bin/mimencode* with the correct encoder.

The *openssl* from the OpenSSL package is installed on most desktop systems. There are two variants of *mimencode*; the one from the *metamail* package requires the *-u* option in the command line. If in doubt, check out the manpage and experiment at the command line.

The mail filter program will add a prefix to the subject line in the **header** of mails classified as spam, ham or unsure. KMail does not need this information as the mail filter program also adds them at another position in the header. You can enable *:spam_flag_subject_string: //*, *:unsure_flag_subject_string: //*, and possibly *:good_flag_subject_string: //* if you prefer an unchanged subject line.

By default, the mail filter program will log messages it processes in *allmail.txt*.

Listing 1: kmail.antispamrc

```
[Spamtool #11]
Ident=crm114
Version=1
Priority=65
VisibleName=CRM114
Executeable=crm -v | grep "CRM114"
URL=http://crm114.sourceforge.net
PipeFilterName=CRM114 Check
PipeCmdDetect=crm -u $HOME/.crm114 mailreaver.crm
ExecCmdSpam=crm -u $HOME/.crm114 mailreaver.crm --spam
ExecCmdHam=crm -u $HOME/.crm114 mailreaver.crm --good
DetectionHeader=X-CRM114-Status
DetectionPattern=SPAM
DetectionPattern2=UNSURE
DetectionOnly=0
UseRegExp=0
SupportsBayes=1
SupportsUnsure=1
```

GLOSSARY

POP3: A mail protocol where the mail client picks up messages from the server and then typically deletes the messages on the server.

IMAP: A mail protocol where the mail client directly accesses messages on the server. KMail supports local email caching, disconnected IMAP (DIMAP) or cached IMAP.

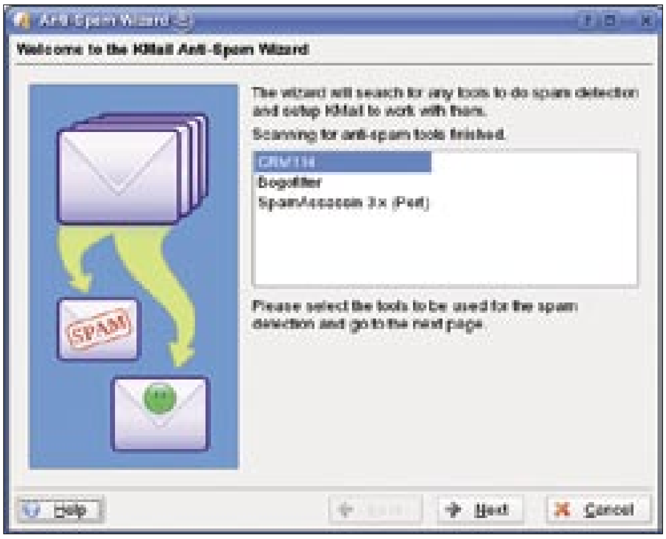
Upstream version: A released developer version of the program.

Static linking: The linker uses function libraries to link the program. In the case of static linking, the required functions are stored directly in the executable. After compiling, a dynamic linker will simply set up a non-permanent link when a program is launched.

Ham: In this context, ham refers to legitimate email messages.

Hash: A unique, short ID for an arbitrarily long data fragment, which accelerates data searching.

Header: The email header, in this case, which includes metadata such as the sender, target or the mail program used. The content of the email is stored in the body.



Anti-Spam Wizard

Welcome to the KMail Anti-Spam Wizard

The wizard will search for any tools to do spam detection and setup KMail to work with them.

Scanning for anti-spam tools finished.

CRM114
Bogofilter
SpamAssassin 3.x (Perl)

Please select the tools to be used for the spam detection and go to the next page.

Figure 1: The KMail antispam wizard provides a convenient GUI for setting up the filter.

If the mail filter is working to your satisfaction, you can save space by disabling logging:

```
:log_to_allmail.txt: /no/
```

The configuration file in the CRM114 package for Debian contained a bug prior to version *20060704a-3* [8]. The path to *mailtrainer.crm* is incorrect. Make sure it is spelled as follows:

```
:trainer_invoke_command: ➘
 /.\/mailtrainer.crm/
```

The commands:

```
cssutil -rb spam.css
```

and:

```
cssutil -rb nonspam.css
```

let you create hash files for identifying messages as spam or ham.

The command creates the specified CSS file if the file does not already exist, and then the command outputs statistics.

The mail filter program accesses two list files to implement a combined blacklist/whitelist or to rewrite mail addresses.

For the program launch, you can create empty files by entering:

```
touch rewrites mfp priolist.mfp
```

(See "Rewrite Rules, and Black/Whitelists" box.)

Prior to KMail integration, it makes sense to test your mail filter installation at the command line. Enter *crm mailreaver.crm* to launch the CRM program, enter some arbitrary text and press Enter followed by [Ctrl] + [D]. The CRM program should now display a couple of header lines that start with *X-CRM114*.

Antispam Wizard

KMail has an antispam wizard that detects any spam filters you install and creates matching filter rules. In our lab, we used KMail from KDE 3.5.5 on Debian Etch and KMail from KDE 3.5.1 on Suse 10.1. See Figure 1.

The KMail Antispam wizard from KDE 3.5.5 does not support CRM114. KDE version 3.5.6 should have a KMail version with CRM114 support [9]. You can teach CRM114 to older versions of KMail by adding an entry for CRM114 to the *kmail.antispamrc* configuration file (see Listing 1).

On Debian and Ubuntu, the file is located in */etc/kde3*; users with Suse must check the */opt/kde3/share/config* directory. As an alternative, you can copy the file to *~/.kde/share/config* and customize it at that location.

You may need root privileges to modify the file. Open the file in a text editor and increase the number of spam filters in *tools=* below *[General]* by one. Then add an entry at the end of the file, and in the first line of the entry, modify the *Spamtool #11* entry to match the number of the entry.

Create a directory for spam and one for unsure messages. To do so, right click *Trash* and select *New subfolder ...*. The antispam wizard does not allow you to select folders for online IMAP accounts; you will need to use a local folder, which is a faster option anyway.

It makes sense to store spam temporarily in a separate folder, instead of in trash, since KMail deletes messages in the trashcan when you quit the program. This gives you a second chance to review spam for false positives.

The antispam wizard is called by *Tools | Anti-Spam Wizard ...* to create matching filter rules in two steps. To do so, first select CRM114 as your spam filter. The second step is to choose the folder for spam, and one for unsure messages.

After completing these steps, KMail is now ready to use CRM114 as a spam filter. The five new filter rules should be at

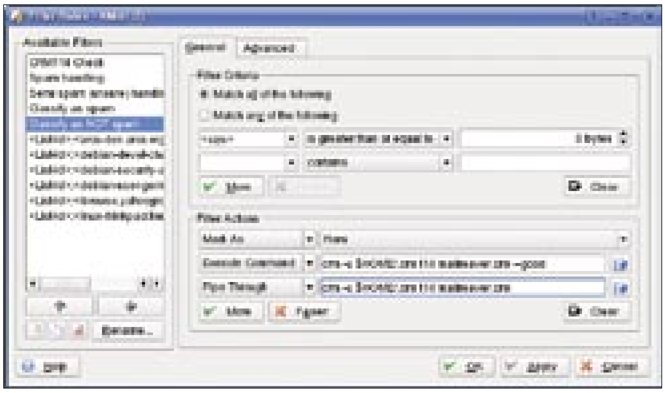


Figure 2: The filter rules for the spam filter need to be at the top of the list.

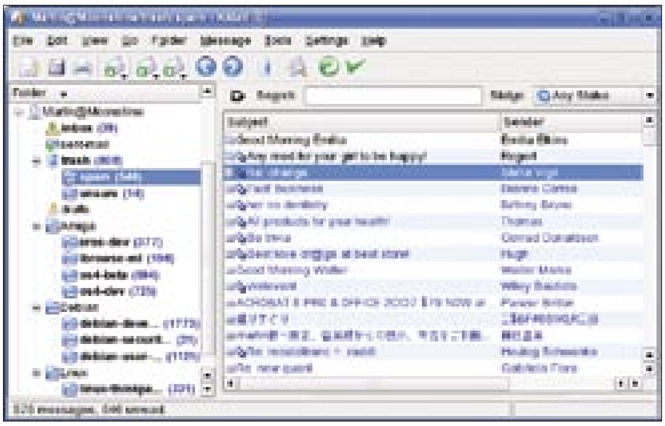
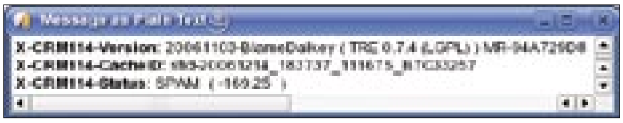


Figure 3: After just a couple of hours of training, CRM will remove most spam.



Message as Plain Text

```
X-CRM114-Version: 20061103-BlameDalkey ( TRE 0.7.4 (LGPL) ) MR-94A72908
X-CRM114-CacheID: sfid-20061218_183737_111675_8TC33257
X-CRM114-Status: SPAM ( -169.25 )
```

Figure 4: Spam all the way as CRM114 says in the X-CRM114-Status header line.

the top of your filter ruleset to make sure that they are always applied. Select:

```
Preferences | ↵
 Configure filter...
```

and use the arrow icons to move the spam filter rules up in the list (Figure 2).

You will see two new icons in the toolbar. The green checkmark lets you flag legitimate mail as such, and the icon that shows yellow with green dots lets you flag unsolicited mail.

KMail will filter ham mail based on any other filter rules you set up or just leave it in your mailbox.

Spam or unsure mail is moved to the folders specified.

Don't try to teach CRM114 with just any old mail. CRM114 is very much like an attentive child at first and it learns quickly, but it is only interested in false positives or negatives, or in messages that it was unable to classify.

Make sure you teach CRM114 regularly – once an hour, if possible – or download new messages bit by bit. You can interrupt the download to do this. The more often you teach CRM114, the less messages you need to tag, as CRM will already have seen them. This makes CRM all the more effective. After just a couple of hours training, CRM successfully removed the lion's share of spam in our lab (see Figure 3).

Watch out for false positives (messages that are incorrectly identified as spam) at first and make sure you tag them as legitimate. Again, CRM114 will learn very quickly.

If you want to refilter ham in the folder for unsure mail to move it to the correct folder, select *Apply filter* | *Apply all filters* from the context menu, or just press [Ctrl] + [J].

For more information on mail classification, you can press [V] to view the message, including the headers (see Figure 4). Look for the *X-CRM114-Status* header line. KMail does not immediately change these header lines with the antispam wizard's filter rules when you tag a mail as ham or spam. Thus, a mail you tag as ham will still have an *UNSURE* or *SPAM* line.

To change this, you need to refilter the mail to tell KMail to reapply the spam filter. Add the filter rule *Run through program* to your *Classify as non-SPAM* and *Classify as SPAM Filter actions* by running the following:

```
crm -u $HOME/.crm114 ↵
 mailreaver.crm
```

See Figure 2.

Conclusions

You can see that as of this writing, installing CRM114 and integrating the program with KMail is still fairly complex. However, it is definitely worthwhile, as you will notice when the volume of spam starts to drop. So let's just say: "Welcome, my name is CRM114, Content Regulator for Mail!" ■

Building CRM114

Download the source code archive [10]; unpack the archive, and change to the new directory. Start by building the *TRE* library, which handles the regular expressions. To do so, *cd* to the directory (*tre-0.7.4*, for example). Configure the sources by running *./configure* or *./configure --enable-static* if you want CRM114 with static linking. Run *make* to compile the source code and then enter *su -c "make install"* to install the library.

You can ignore error messages referring to tests. The *ls -l /usr/local/lib/libtre** should list the newly installed library files. If you need dynamic linking, run *ldconfig* as root. Your */etc/ld.so.conf* file needs a path entry for */usr/local/lib*, in this case.

Then move up a level in the directory tree by entering *cd ..*. The CRM files are installed to */usr...*, by default. You can modify the *prefix* variable if you want to install in */usr/local*. For dynamic linking, add a pound sign at the start of the *LDFLAGS += -static -static-libgcc* line. Then run the *make clean*, *make*, and *su -c "make install* commands, in that order.

Rewrite Rules, and Black/Whitelists

rewrites.mfp tells CRM114 to replace your name, your own email address, and (if needed) the mail gateway IP address and name with generic names (see Listing 2). You first need to specify what to replace, followed by >-> and the replacement string *Search>->Replace*. >--> supports searching over multiple lines. Regular expressions are also supported.

This is useful for protecting your privacy. At the same time, it reduces the learning curve for CRM114 as it removes your details, which are not important for the spam/ham decision. In our lab, CRM114 proved to be extremely accurate and efficient, even without rewrites.

There is typically no need for a blacklist or whitelist as CRM114 is a very fast learner. The new *mailreaver.crm* simply uses a combined black/whitelist, *priolist.mfp*. Debian and Ubuntu have an example at */usr/share/doc/crm114/examples*, and there is one in the Upstream archive. This removes the need for manual editing whitelists and blacklists, which is always prone to error.

Listing 2: rewrites.mfp

```
Yourname@your.address.org>
   ->MyMailAddress
[[:space:]]Your Name>-> MyName
```

INFO

[1] Patch for KMail: *http://commit-digest.org/issues/2006-11-05/moreinfo/600586/*

[2] CRM114 website: *http://crm114.sourceforge.net/*

[3] CRM114 and CRM114 mail filter FAQ: *http://crm114.sourceforge.net/FAQ.txt*

[4] CRM 114 mail filter HOWTO: *http://crm114.sourceforge.net/CRM114_Mailfilter_HOWTO.txt*

[5] "CRM114-Training via SquirrelMail" by Steve Pellegrin: *http://www.convoglio.com/crm114/*

[6] Backports.org: *http://www.backports.org*

[7] Current version of BlameDalkey: *http://crm114.sourceforge.net/crm114-20061103-BlameDalkey.i386.tar.gz*

[8] Debian bug report: *http://bugs.debian.org/394476*

[9] KDE wishes and patch: *http://bugs.kde.org/show_bug.cgi?id=136261*

[10] Current version of BlameDalkey: *http://crm114.sourceforge.net/crm114-20061103-BlameDalkey.src.tar.gz*