

By Sean

Published: 2008-07-22 14:03

How To Block WebPages Based On Keywords Or Phrases With SafeSquid Proxy Server

Keyword Filtering allows you to block web pages, depending on the words and phrases found in the page's title, meta tags and body. Keyword filtering in [SafeSquid](#) uses a 'weighedkeyword scoring' method. It analyzes web pages, and searches for specified, unacceptable words or phrases. Every time a word or phrase is found, it assigns the score associated with the word or phrase to the web page. When the total score of the web page exceeds a threshold limit, it is blocked. The advantage of using Keyword filtering, is that it can block web pages containing specific content, like pornography, without depending on any database, like URL Blacklist. Also, it will block a web page, even if it is being accessed through an anonymous proxy, provided it is not an HTTPS connection.

SafeSquid allows the use of regular expressions for defining keywords. Use of regular expressions, allows you to precisely define your search expressions, and also differentiate between words like 'sex' and 'sensex'.

The weighed keyword scoring method ensures that a web page is not unnecessarily blocked, if it contains a single unacceptable word. For example, if you blocked all web pages that contain the word 'sex', it will also block web pages that require you to fill a form, with the option - 'Sex: Male / Female'. The correct method is to block a page, if it contains a combination of unacceptable words, each assigned a specific score.

The first step is to generate a list of keywords that are generally found in web pages that you would like to block. I found Google's 'Keyword Tool', to be extremely helpful and easy to use (<https://adwords.google.com/select/KeywordToolExternal>). It allows you to generate a list of keywords, related to the keyword that you are looking for. Additionally, it has a 'Website Content Option'. It lets you enter the URL of a website, scans the page and then lists the keywords used in the page, and suggests relevant keywords. The list is divided into several groups of similar keywords. You can then export and save the list to a text or CSV file.

Once you are ready with your keywords list, you can use it in SafeSquid *Keyword Filter* section. To access the *Keyword filter* section, open the SafeSquid Web Interface and go to *Config => Keyword filter*.

Option	Value
Enabled	Yes: No:
Threshold	
Template	

Add

Verify that the section is enabled (*Enabled=Yes*). The default *Threshold* is *100*. Now, if your list consists of 5 groups of keywords, and you would like to block a page, if it matches a word or phrase from all the 5 rules, then you can specify a score of 20 to each rule, to reach the *Threshold* value (5 X 20 = 100). Similarly, if you would like to reach the *Threshold* value, if 4 rules match, then give each rule a score of 25.

Click on *Add* under the *Keyword* subsection, to add a new rule, as shown below:

Option	Value
Enabled	true
Comment	Keywords for blocking 'XYZ' related websites - Rule 1/7
Profiles	!main-sites
Mime type	text/html
Keyword	b(phrase one phrase two phrase three phrase four phrase five)b
Score	25

Note *!main-sites* in the *Profiles* field. *main-sites*, as explained in previous howtos, is a profile created under the *Profiles* section. It applies the profile *main-sites* to specific websites, that you do not want to be effected by any filter. The '!' before main-sites denotes NOT, meaning apply this rule to all websites, but NOT, or EXCEPT, the ones with main-sites profile.

Also note the entry in the *Keyword* field. 'b' means a blank space and the separator '|' means 'OR'. So, the rule means - a blank space, followed by phrase one OR phrase two OR phrase three OR phrase four OR phrase five, followed by a blank space.

This rule will apply a score of 25 to a web page, if it contains any of the phrases specified in the Keyword field.

Similarly add other rules and specify their scores. After you have added all the rules, try accessing a web page that you know contains the specified phrases, and check if it gets blocked. To check how much a web page scores, use the special SafeSquid URL Command *xx--score*, like this -

`http://xx--score.www.somewebsite.com`

This will display the score of the web page. Now, depending on the score, either fine tune your rules, score for each rule, or the *Threshold* value.

SafeSquid has predefined sample rules for blocking porn and proxy websites. You can also download these sample rules from the website (use your SafeSquid forum ID to authenticate)-

Porn keywords - http://downloads.safesquid.net/free/general/sample_rules/porn_keywords.xml

Anonymous proxy - http://downloads.safesquid.net/free/general/sample_rules/anonproxy.xml

To import a rule into your existing SafeSquid config file, copy the downloaded file to the SafeSquid server, open the interface and click on *Load settings* in the *Top menu*. Specify the full path for the file in the *Filename* field (e.g. /usr/local/src/porn_keywords.xml), select *Overwrite = No* and click on *Submit*. This will append the rule snippet into your existing *Keywordfilter* section.

You can also directly import the file into your config.xml, by specifying the above mentioned URL in the *Filename* field, with your username and password, like this -

`http://username:password@downloads.safesquid.net/free/general/sample_rules/porn_keywords.xml`

Remember to set *Overwrite* to *No*, before you click on *submit*. Check if the rules have been imported. In case of any problems, just restart SafeSquid

service, and your original config.xml will be loaded. If everything is fine, then click on *Savesettings* in the *Top menu* to make the changes permanent.

Also see:

- [Deploying A Content Filtering Proxy Server To Distribute Controlled Internet Access With SafeSquid](#)
- [Set Up Gateway Level Virus Security With ClamAV And SafeSquid Proxy](#)
- [How To Set Up Internet Access Control And Internet Filtering With SafeSquid Proxy Server](#)
- [How To Control Access To Unwanted Websites Using URL Blacklist With SafeSquid Proxy Server](#)
- [How To Configure Granular Bandwidth Management Rules In SafeSquid Proxy Server](#)
- [How To Control Download Of Files And Mime Types In SafeSquid Proxy Server](#)
- [How To Block Ads And Banners In SafeSquid Proxy Server](#)
- [How To Block Cookies From Unwanted Websites With SafeSquid Proxy Server](#)
- [Enhance Security By Removing ActiveX Control Codes From Web Pages With SafeSquid Proxy Server](#)