SFproxy

An Indexing HTTP Proxy
Edition 0.9, for SFproxy 0.9
July 1995

by Kai Großjohann and Ulrich Pfeifer

1 Overview

SFproxy has several modes of operation. On the one hand, it can be an HTTP proxy, like many other HTTP proxies. As such, it performs a different function than other proxies: It watches requests and responses pass by and if a request is a `GET` and if a document of content type `text/html` is sent as response, that document is indexed under the URL given in the request. This way, you get a "better hotlist" because you can search it.

Another mode of operation is that you can create a searchable index from a list of URLs. One possible form of a list of URLs is the '`.mosaic-global-history`' file maintained by the `Mosaic` WWW-Browser.

2 General Information

Here is some information on how SFproxy works in its two modes.

2.1 Proxy Mode

There are two ways of using SFproxy in proxy mode. The first way is to start SFproxy in server mode. This corresponds to the `-server` option. In this mode, SFproxy runs as a background process, listening for connections on a specific port. When a WWW browser (more generally, a client) connects to that port, SFproxy then forks a child which reads a request from the client, passes it on to the appropriate server and then gets the response from the server and passes it on to the client. If the request was a `GET` and if the server responded with a document, that document is indexed with WAIS under the URL indicated by the request.

You can then use any WAIS client (preferably `SFgate`) to query this WAIS database.

The other way of running SFproxy in proxy mode is to use the daemon mode which corresponds to the `-daemon` option. In this mode, the `inetd` program takes over the task of listening on the port and of forking off and instance of SFproxy. An appropriate entry must be made in the `inetd` configuration file for this to work.

2.2 List Processing Mode

In this mode, instead of waiting for HTTP requests, SFproxy reads a file of URLs (for example, the Mosaic global history file), creates a request on its own, and indexes the corresponding document returned by the server.

2.3 More Details

SFproxy only understands about documents of content type `text/html`. Other documents are simply passed through in proxy mode and discarded in list processing mode, respectively.

SFproxy only indexes the response from a server if the status code of the response indicates success.

SFproxy understands HTTP/1.0 only. (HTTP/0.9 requests and responses are passed through but not processed any further.)

SFproxy does not index the same URL twice, ie changes in the documents do not propagate to the WAIS database. A workaround is the `-recreate` option which discards the whole database and re-fetches all of the documents contained therein.

3 Invokation

3.1 Options selecting the mode

SFproxy can be in a number of modes, the most important are *daemon mode*, *server mode* and *list processing mode*. Less important modes are *recreate mode* and *printurls mode*. Here's which option invokes which mode.

Symbols in upper case indicate the type of the argument required. `NUM` means a number, `STR` means a string. Square brackets are used if the argument is optional.

`-server NUM`
: This option invokes server mode. SFproxy listens on the given port for connections. When a connection is received on this port, a child is forked to process the request.

`-daemon`
: This option invokes daemon mode. A request is read from `STDIN` and processed. The answer goes to `STDOUT`.

`-list STR`
`-momspider STR`
`-mosaichotlist STR`
`-netscapehotlist STR`
: These options invoke list processing mode. The URLs are read from the given file. '`-`' means use `STDIN`. `-list` is a general option, whereas `-momspider`, `-mosaichotlist`, and `-netscapehotlist` are tailored for specific list formats.

`-addurl STR`
: This option adds a single URL to the database. The URL is the command line argument.

`-recreate`
: Use this option if you already have a database and you want to bring it up to date. Please note that in server, daemon, and list processing mode, no URL will be indexed twice, ie the database may not be up to date anymore after a while. Recreate mode refetches all of the documents in a database and indexes them.

`-printurls`
: Print the URLs stored in the WAIS database. (Use the other options to specify the database.)

3.2 Other options

Here's the list of options, together with their meanings. Please note that for some options, a default value is given. Your installation may have a different default value, depending on your configuration. See Chapter 4 [Configuration], page 6 for more information.

Symbols in upper case indicate the type of the argument required. `NUM` means a number, `STR` means a string. Square brackets are used if the argument is optional.

`-debug`
: This option prints debugging output to `STDERR` (currently).

`-ddebug`
: This option prints a **lot** of debugging output to `STDERR` (currently).

`-lockwait NUM`
: When SFproxy tries to get a lock on a file and that fails, it waits this number of seconds before trying again. If the given number of seconds is negative (say, -5), the number of seconds really waited derives from this number and the last digit of the current process id (in the example, 5 plus the last digit of the pid).

`-lockexpire NUM`
: When a lock is older than this number of days, it is considered stale and broken.

`-nice NUM`
: In list processing mode, a child is forked to process each specific request. This option says not to use more than the given percentage of system resources, in particular number of file descriptors, number of inodes, number of process table entries, amount of swap space. Forking a child is delayed until all of these are below the given threshold.

  The default value is 80 percent.

`-nicewait NUM`
: If the option `-nice` is used, this option gives the number of seconds to wait between two checkings of resource usage.

`-maxchildren NUM`
: In list processing mode, at most this number of children may be running concurrently. The default is no limit for the number of children.

`-dir STR`
: Before doing anything, SFproxy `chdir`s to the directory given here. Ie this is the directory the database resides in. If this option is not used, the default value of `.` is used.

`-database STR`
: This gives the name of the WAIS database. This may not contain any slashes. Use the `-dir` option to specify a directory. If this option is not given, the default value of `SFproxy-db` is used.

`-urlfile STR`
: The URLs stored in the WAIS database are kept track of in a dbm file. This option gives the name of that file, if it differs from the default value of `Z-url`, where `Z` is the name of the database.

`-indexprefix STR`
: Files to be indexed by WAIS are written to a temporary file. Its name is the concatenation of the following values: index prefix, database name, the string ".", and the current process number (in this order).

`-waisindex STR`
: This is the name of the `waisindex` program to be used, ie the complete path. If this option is not given, the default value '`/usr/local/ls6/wais/bin/waisindex`' is used.

`-proxy STR`
: This is the name of the HTTP proxy host, if one is to be used. Default is the value of the http_proxy environment variable.

`-proxyport STR`
: This is the port on the proxy host to be used, if a proxy is used at all. Default is "http".

`-noproxy [STR]`
: If the optional string is not given, no HTTP proxy is used at all. If the optional string is given, the HTTP proxy is used only for hostnames not matching this regexp.

`-re STR`
: If the `-list` option is used, a regexp may be given. Each line of the file of URLs is matched against this regexp.

  If the line does not match this regexp, it is skipped.

  If the line matches this regexp, the URL is that part of the line that matched this regexp. But see the option `-reindex`, as well.

`-reindex NUM`
: If this option is used, the regexp given with `-re` must contain at least NUM plus one pairs of parentheses. The URL is only that part of the line that matches the part of the regexp between this pair of parentheses. 0 means first pair of parentheses, 1 means second, and so on. For example, if you have a file where each line contains a three digit number, then a space, and then the URL, you could use the following combination of options (please note the use of single quotes to escape the regexp from the shell):

```
-re '[0-9][0-9][0-9] (.*)' -reindex 0
```

4 Configuration

At the beginning of the file SFproxy you will find a section titled “Configuration Variables”. Below that, there is a section titled “Other Variables”.

The “Configuration Variables” section contains a number of variables, together with their default values. When installing SFproxy at your site, you might want to change some of these values so that they make more sense for you. The defaults given in this section may be overridden with command line options.

You probably won’t need to change any of the “Other Variables”. You might want to look at the `$debug`, `$ddebug` and `$debug_fh` variables. I don’t understand the `$sockaddr` variable, either ;–)