

Filtry treści internetowych Privoxy i Webcleaner

# STRAŻNICY

Filtry treści chronią prywatność użytkownika sieci WWW

i zabezpieczają go przed zalewem niepożądanych reklam.

W tym artykule zademonstruję dwie aplikacje filtrów treści internetowych na licencjach Open Source.

THOMAS LEICHTENSTERN

**F**iltry treści są użyteczne do zabezpieczenia przed niekontrolowanym przepływem danych pomiędzy przeglądarką a serwerem WWW. Dobry filtr treści pozwala na przepływ tych danych, które są rzeczywiście pożądane przez użytkownika, blokując reklamy, robaki internetowe, ciasteczka i niechciany kod Javascript. Niektóre filtry potrafią również zarządzać ruchem wychodzącym. Prawidłowo skonfigurowany filtr treści może zabezpieczyć użytkownika nawet przed konsekwencjami dziur w przeglądarkach WWW, które niestety są nadal zjawiskiem bardzo powszechnym.

Ten artykuł omawia filtry treści Privoxy [1] oraz Webcleaner [2]. Obydwa narzędzia obsługują funkcje filtrowania treści, lecz Privoxy skupia się na treści WWW, Webcleaner posiada zaś kilka dodatkowych funkcji, jak filtr przeciwwirusowy czy kompresor obrazów.

## Przekierowanie dostępu

Filtry treści działają jak serwery pośredniczące (*proxy*), to znaczy są pośrednikami pomiędzy przeglądarką a serwerem WWW. Aby skorzystać z filtra treści należy przekierować przeglądarkę na adres, na którym działa filtr treści, modyfikując odpowiednio ustawienia połączenia w przeglądarce. Jeśli filtr działa w systemie lokalnym, należy wpisać adres 127.0.0.1 („localhost”).

## Privoxy

Privacy Enhancing Proxy - lub w skrócie Privoxy - jest oparty na programie Junkbuster. To prosty filtr treści nie posiadający funkcji buforowania. Privoxy, w odróżnieniu od zwy-

kłego filtra adresów URL, kontroluje wymianę danych z serwisem WWW w oparciu o reguły dotyczące treści stron.

## Instalacja

Privoxy jest dostępny dla większości dystrybucji, nie powinno być więc problemu z odnalezieniem odpowiedniego pakietu DEB lub RPM. Użytkownicy Debiana mogą skorzystać z repozytorium *Universe* i zainstalować program za pomocą polecenia `apt-get install privoxy`. W Suse 9.3 Privoxy można zainstalować bezpośrednio z nośników instalacyjnych za pomocą programu Yast. W tym przypadku program jest instalowany w konfiguracji *chroot*. Z tego powodu konfiguracja i pliki dziennika znajdują się w katalogu `/var/lib/privoxy/`.

## Konfiguracja

W domyślnej instalacji Privoxy nasłuchuje na adresie localhost (127.0.0.1). Jeśli program ma działać w trybie sieciowym z dostępem dla innych maszyn, należy w pliku `/etc/privoxy/config` skonfigurować adres interfejsu sieciowego, na przykład 192.168.0.1. Jeśli adres zostanie pominięty, usługa będzie nasłuchiwać na wszystkich interfejsach sieciowych, co nie jest zalecane, szczególnie w przypadku komputerów z bezpośrednim dostępem do Internetu.

Reguły filtra można ustawić za pomocą interfejsu WWW. Dostęp do interfejsu jest pod adresem `privoxy.org/config`. Najpierw jednak należy skonfigurować Privoxy jako serwer pośredniczący przeglądarki WWW. W Debianie trzeba dodatkowo uaktywnić opcję obsługi interfejsu

WWW. W pliku `/etc/privoxy/config` należy odszukać opcje `enable-remote-toggle` oraz `enable-edit-actions` i ustawić ich wartości na 1. Po dokonaniu zmian trzeba przeładować aplikację za pomocą polecenia `/etc/init.d/privoxy restart`. Privoxy nie obsługuje uwierzytelniania użytkowników, zatem każdy użytkownik z dostępem do filtra treści może zmienić jego konfigurację.

## Filtry

Konfiguracja Privoxy wykorzystuje pliki filtrów oraz pliki akcji. Filtry zawierają reguły, na przykład reguły blokowania bannerów o rozmiarach przekraczających określony zakres. Pliki akcji odwzorowują reguły na adresy. Adresy mogą natomiast być zdefiniowane w postaci zwykłych adresów URL lub wzorców dopasowań reprezentujących fragmenty adresów na przykład serwisów reklamowych. Wpis `ad*.example.com` zostanie dopasowany na przykład do wszystkich poddomen domeny `example.com`, których nazwy rozpoczynają się od znaków `ad`, po których występuje dowolny ciąg znaków.

## Hasta la Vista, Baby

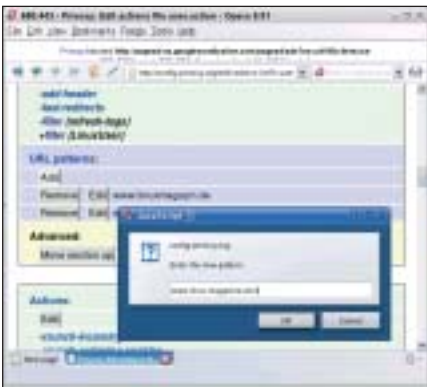
Domyślny plik filtra (`/etc/privoxy/default.filter`) zawiera sporą kolekcję reguł. Użytkownicy niemający dużej wprawy w pracy z wyrażeniami regularnymi, nie powinni jednak próbować modyfikować zawartości tego pli-



Rysunek 1: Główne okno Privoxy oferuje menu i odnośniki do różnych funkcji aplikacji.



Rysunek 2: Privoxy wyświetla reguły filtra w swoim interfejsie WWW.



Rysunek 3: Sposób definiowania akcji jest prosty, użytkownik szybko nauczy się definiować własne reguły.

ku. Privoxy nie udostępnia interfejsu wspomagającego modyfikację plików filtrów; należy w tym celu wykorzystać dowolny edytor tekstów. Poniższy przykład prezentuje składnię typowej definicji reguły filtra:

```
FILTER: LinuxMagazine ↗
Przykładowa reguła filtra
s/rain(?!.com)/sun/ig
```

Słowo kluczowe *FILTER*: definiuje nazwę nowej klasy (*LinuxMagazine*) oraz opis (*Przy-*

*kładowa reguła filtra*). W interfejsie WWW opis reguły filtra można znaleźć w sekcji Privacy. Drugi wiersz zawiera samą regułę. W tym przypadku zastępuje ciągi znaków *rain* ciągami *sun*. Klasa może składać się z dowolnej liczby reguł, które można aktywować za pomocą jednego kliknięcia w interfejsie WWW. W Internecie można znaleźć sporą liczbę gotowych do wykorzystania list reguł filtra [4].

### I ... akcja!

Najbardziej precyzyjnie przygotowana reguła nie jest wiele warta, jeśli nie określi się dla niej celu. Do tego właśnie służą pliki akcji. Plikami akcji są *user.action* oraz *default.action* i obydwa te pliki można edytować za pomocą interfejsu WWW. Adresem tych opcji jest *config.privoxy/show-status*. Oba pliki mają identyczną strukturę, lecz są wykorzystywane do innych celów. Plik *default.action* określa zachowanie domyślne, *user.action* obsługuje określone aplikacje.

Plik *default.action* zawiera domyślne reguły, które są stosowane w przypadku, gdy nie mają zastosowania reguły z pozostałych plików akcji. Privoxy wykorzystuje trzy standardowo zdefiniowane tryby (policy), przydatne szczególnie początkującym użytkownikom. W interfejsie WWW można zatem wybrać tryb od *Cautious* po *Adventuresome*.

Kolejne sekcje pliku definiują poszczególne tryby obsługi adresów, opisując sposób ich obsługi w oparciu o wzorce dopasowań, na przykład *ad\** lub *\*banner\**. Konfiguracja globalna ogranicza się z grubsza do możliwości wyboru jednego z domyślnych trybów.

Reguły użytkownika są definiowane w sekcji *user.action*. Jeśli Privoxy w domyślnej konfiguracji blokuje dostęp do strony, której blokować nie powinien, można wykorzystać odnośnik <http://config.privoxy.org/show-url-info>, pod którym odszukamy filtr odpowiedzialny za takie działanie aplikacji. Klikając przycisk *Insert new Section at top* w sekcji *user.action*, a następnie *Add*, możemy dodać adres URL naszej strony. *Edit* pokaże listę reguł z *default.filter*, które mają zastosowanie w tym konkretnym przypadku. Domyślne ustawienie dla tych reguł to *No Change*. Każda modyfikacja dokonana w tym miejscu będzie miała priorytet nad domyślnymi trybami.

### Prace trwają

Privoxy nie przeszkadza w codziennej pracy z przeglądarką i prawie nie opóźnia ładowania stron WWW, nawet w przypadku dużych dokumentów. Oczywiście, należy pamiętać, że by do obsługi tej aplikacji nie wykorzystywać

całkiem muzealnego sprzętu. Dolną granicą jest procesor 500 MHz oraz 256 MB RAM.

Program wyświetla poprawnie większość stron, nawet w przypadku ustawienia trybu *Cautious*. Jeśli jakiś serwis nie działa, można sprawdzić reguły, które mają zastosowanie do danego adresu, lecz kłopot odnalezienia w gąszczu reguł właśnie tej, która jest odpowiedzialna za błąd w pracy ze stroną WWW, spoczywa na użytkowniku.

Jeśli strona WWW jest całkowicie zablokowana, Privoxy nadal daje szansę uratowania się z tej sytuacji. Służy do tego odnośnik *go there anyway* („idź tam pomimo wszystko”). Kliknięcie go spowoduje tymczasowe zablokowanie wszelkich filtrów dla danego adresu. Program udostępnia tak zwane bookmarklety służące do szybkiego wyłączania i włączania filtrów. Funkcję tę można odnaleźć klikając *Privoxy – Toggle Privoxy* na stronie <http://config.privoxy.org>. Dzięki tej opcji można całkowicie wyłączyć lub ponownie włączyć funkcje filtrujące Privoxy za pomocą jednego kliknięcia.

### Privoxy: podsumowanie

Dużo pochwał należy się twórcom programu za wzorcową wręcz dokumentację opisującą szczegółowo wszystkie jego funkcje. Kombinacja plików definicji filtrów i akcji może początkowo wpędzać w zakłopotanie, lecz przy bliższym spojrzeniu okazuje się być bardzo dobrym pomysłem.

W ostatecznym rozrachunku Privoxy sprawia wrażenie dojrzałego, dobrze przemyślanego narzędzia. W naszym laboratorium pracował poprawnie bez żadnych problemów i w domyślnej konfiguracji dobrze spełniał postawione przed nim zadanie zapewnienia przeglądania stron WWW pozbawionych reklam.

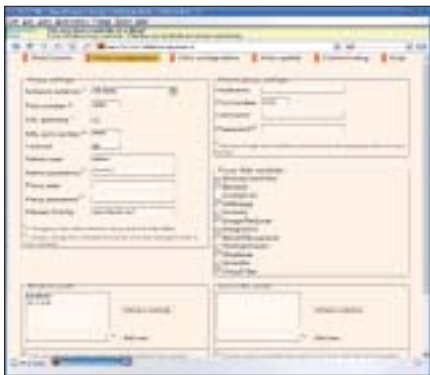
### Webcleaner

Drugi przetestowany przez nas program to Webcleaner, który może pochwalić się wręcz niewiarygodną listą możliwości. Oprócz filtrowania treści, twórcy programu położyli duży nacisk na możliwości dodatkowe, jak kompresja i skalowanie obrazów, filtrowanie wirusów i wykrywanie i poprawianie „w locie” błędów w formacie HTML.

### Instalacja

Webcleaner był tworzony głównie z myślą o Debianie, więc najprościej jest zainstalować go na tej dystrybucji lub pokrewnej, na przykład Ubuntu. Program może być zainstalowany również na innych dystrybucjach, w tym również na Suse.

Webcleaner do działania wymaga programu *runit* oraz interpretera Pythona w wersji



Rysunek 4: Okno interfejsu konfiguracyjnego Webcleanera.

2.4, włączając w to pakiety deweloperskie tego języka. Aby skompilować Webcleanera z kodu źródłowego potrzebny jest również kompilator języka C, na przykład *gcc*.

Użytkownicy zainteresowani pełnym repertuarem możliwości Webcleanera muszą mieć zainstalowane następujące narzędzia i biblioteki:

- PIL (Python Image Libraries) – niezbędne do kompresji i skalowania obrazów;
- Open-SSL i Python-openssl – w celu umożliwienia obsługi połączeń szyfrowanych SSL;
- Clamav (*clamd*) – do obsługi filtra antywirusowego.

Jako dodatek można również zainstalować bibliotekę *psyco*. Zgodnie z twierdzeniem twórców Webcleanera *psyco* przyspiesza wydajność skryptów Pythona 2 do 100 razy.

Użytkownicy Debiana mogą zainstalować Webcleanera za pomocą *apt-get* lub *Synaptic*a. Użytkownicy Suse 9.3 powinni ręcznie skonfigurować *runit*, *psyco*, oraz PIL, pozostałe pakiety powinny być dostępne na nośnikach instalacyjnych Suse.

## Instalacja ze źródeł

Rozpakowujemy archiwum ze źródłami programu: *tar xfvz webcleaner-2.29.tar.gz*, po czym przechodzimy do utworzonego w ten sposób katalogu. Następnie wykonujemy standardową kombinację poleceń *./configure && make*. Skrypt może zakończyć się błędem z powodu brakującej biblioteki (*/lib/cpp*), dotyczy to przede wszystkim Debiana. W takim przypadku należy zainstalować pakiet *opencc++* i ponownie wywołać polecenie.

Po zakończeniu kompilacji należy skompilować pliki Pythona wywołując polecenie *python setup.py build*. Wywołanie *python setup.py install* zainstaluje Webcleaner w systemie.

Jeśli ktoś planuje wykorzystywać funkcję pośrednika WWW w zaszyfrowanych połączeniach SSL, należy zainstalować niezbęd-

ne certyfikaty. Służy do tego polecenie *webcleaner-certificates install*. Na koniec wywołujemy polecenie *make installservice* konfigurujące demona Webcleaner, który jest monitorowany przez *runit*, dzięki czemu jest uruchamiany natychmiast po wywołaniu polecenia. W Suse przed uruchomieniem skryptu należy utworzyć katalog */var/service/*.

## Konfiguracja

Webcleaner może być dostępny bezpośrednio lub jako nadrzędny serwer pośredniczący Squida. W takim przypadku przeglądarka łączy się ze Squidem, skąd następuje przekazanie połączenia do Webcleanera, a stamtąd do Internetu. Aby skorzystać z takiej konfiguracji należy w przeglądarce jako port serwera pośredniczącego skonfigurować port 3128. Jeśli dostęp za pośrednictwem Webcleanera ma odbywać się z innych maszyn, należy dopisać do konfiguracji następujące opcje:

```
....
060 acl localnet src 2
192.168.0.0/255.255.0.0
....
097 http_access allow localnet
....
```

Konfiguracja bezpośredniego wykorzystania Webcleanera (bez Squida) wykorzystuje port 8080, który należy skonfigurować w przeglądarce w ustawieniach serwera pośredniczącego. Webcleaner jest konfigurowany za pośrednictwem interfejsu WWW dostępnego pod adresem *http://127.0.0.1:8080*. Przed pierwszym uruchomieniem Webcleanera należy skonfigurować hasło dostępu administratora. W tym celu w pliku */usr/share/webcleaner/config/webcleaner.config* w opcji *adminpass=* należy wpisać skrót MD5 hasła, po czym ponownie uruchomić Webcleanera poleceniem *kill -HUP ID-processu*. Można już skorzystać z interfejsu WWW logując się jako *admin* z wykorzystaniem skonfigurowanego przed chwilą hasła.

Sekcja *Proxy Configuration* zawiera podstawowe ustawienia programu. Dostępne moduły filtrów można sprawdzić w *Proxy filter modules*. Oto najważniejsze z nich:

- *Blocker* – filtr adresów URL;
- *compress* – kompresja przesyłanych plików;
- *header* – modyfikacja lub usuwanie wybranych pól nagłówka HTTP;
- *Image Reducer* – kompresja obrazów za pomocą ustalonego poziomu kompresji formatu JPEG.
- *Rewriter* – analiza i modyfikacja kodu HTML i Javascript.

U dołu okna pod pozycją *Allowed Hosts*, można skonfigurować adresy maszyn, które będą użytkownikami aplikacji. W oknie *Don't filter Hosts* wpisuje się natomiast maszyny, które mają mieć dostęp do aplikacji, ale których treści nie mają być filtrowane.

Sekcja *Filter configurations* zawiera reguły dla plików. Wpisy w lewej kolumnie określają nazwy katalogów, a reguły dla tych katalogów są wyświetlane po kliknięciu odpowiedniego katalogu w środkowej kolumnie. Kliknięcie reguły natomiast spowoduje otwarcie menu konfiguracyjnego w prawej kolumnie. Aby stworzyć nową regułę należy kliknąć *New rule*. Spowoduje to otwarcie odpowiedniego menu konfiguracyjnego w prawej kolumnie, gdzie można wybrać odpowiednią akcję.

Sekcja *Content rating* służy do kontroli ocen stron WWW. W naszych testach Webcleaner sporadycznie reagował na nasze próby modyfikacji tych ustawień.

## Duże ambicje

W naszym laboratorium Webcleaner ujawnił kilka poważnych słabości. Po włączeniu filtra antywirusowego niektóre transmisje nie działały w ogóle lub działały niepoprawnie. Po otwarciu spreparowanej strony WWW zawierającej świeży szkodliwy kod wykorzystujący dziurę w przeglądarce, komputer testowy zawiesił się. Brama SSL nie chciała działać w naszych testach. Gdy próbowaliśmy uzyskać dostęp do zaszyfrowanego adresu URL, przeglądarka wyświetlała pustą stronę. Interfejs WWW wydaje się w nieprzewidziany sposób akceptować niektóre ustawienia, a odrzucać inne.

Dokumentacja Webcleanera nie zawiera żadnej pomocy do takich problemów. Nie można w niej znaleźć również żadnych wskazówek dotyczących mechanizmu działania programu. W połączeniu z bardzo nieintuicyjną konfiguracją filtrów wrażenia z pracy z programem są niezbyt pozytywne. Nasze próby skomunikowania się z twórcą projektu pozostały bez echa, aż do momentu publikacji artykułu.

Jedyna oczywista zaleta stawiająca Webcleanera nad Privoxy polega na obsłudze uwierzytelniania, zarówno na potrzeby administracyjne, jak i samego korzystania z pośrednika. ■

## Dodatkowe informacje

- [1] Privoxy <http://www.privoxy.org>
- [2] Webcleaner <http://webcleaner.sourceforge.net>
- [3] Junkbuster <http://internet.junkbuster.com/>
- [4] Reguły Privoxy <http://www.neilvandyke.org/privoxy-rules/>