LANCOM Techpaper
Content Filter

The architecture

Content filters can be implemented in a variety of different architectures:

- Client-based solutions, where filter software is installed directly on a desktop computer, are suitable for private requirements. These applications, often called "parental control", are configured in such a way that they check target web sites against a central database server, and they allow or prohibit access accordingly.
- In very large organizations it is worthwhile operating dedicated database servers that allow filter rules to be customized in an individual and specific manner. These solutions demand considerable financial and organizational resources due to the effort of running backup servers and maintaining the rules on a regular basis.
- An integrated solution, where the content filter is implemented on the central gateway, is suitable for medium- to large-sized companies. The LANCOM Content Filter is one example of this type of integrated solution where the filter uses the functions of an existing firewall. One major advantage over the client-based solutions is that the content filter can be configured centrally, meaning that all workstations can be monitored at any time in accordance with a standard policy with little effort. The LANCOM Content Filter consists of a number of components:
  - Timed control for the selection of policies
  - Firewall for the selection of the data streams that are to be examined
  - Content filter that acts as a proxy, accepting the data to be examined and performing the desired actions
  - Rating server that examines the Internet sites and checks to see if they belong to specific categories

The filter concept

The LANCOM Content Filter functions by combining configured rules and the rating server from IBM Tivoli Security Solutions. At the technical level, the content filter inspects each HTTP request according to the following criteria:

- Who is attempting to access the Internet? The answer to this question is supplied by firewall rules that identify specific users or user groups via IP or MAC addresses.
- When is access allowed? This is where the timeframes apply that are used to manage the validity of the filter rules.
- Which URL should be opened? The content filter checks whether the web site belongs to a blacklist, a whitelist, and to one or more categories that are activated via user-defined profiles for a user group.
- What action should the content filter perform? The content filter rules can permit or prohibit access to specific web pages or allow restricted access only.

The first two items are configured by defining appropriate firewall rules and timeframes. In order to check the current URL, the content filter first makes use of the configured blacklists or whitelists. If the URL is found in the blacklist, it is blocked without any further checking. If the URL is found in the whitelist, it is permitted without any further checking. The whitelist has priority over any possible conflicting blacklist entry. When defining blacklists and whitelists, different paths can be entered separately for the same domain

LANCOM Techpaper
Content Filter

meaning that certain areas of a URL can be rated differently.

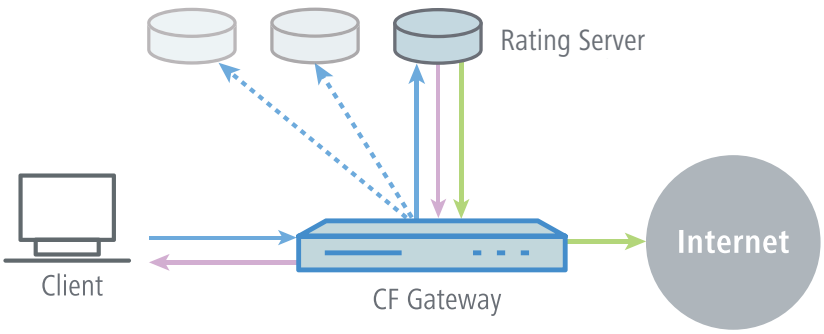
If the URL check based on blacklists and whitelists returns no result, the content filter will forward the query to a rating server. Data is anonymized before being processed so that no inference can be drawn as to the user's identity. The rating server examines the requested URL against its database and delivers the result via the LANCOM Content Filter to the client's browser.

Entries in the database primarily contain the stored URL and the category that the URL has been assigned to. One URL can be assigned to several categories. Furthermore, individual sub-pages can be categorized in different ways so that, for example, a different categorization is valid for each of the three URLs below:

- www.mycompany.eu
- www.mycompany.eu/news
- www.mycompany.eu/downloads

In the same way, individual hosts within the same domain can be assigned to different categories

- www.mycompany.eu
- downloads.mycompany.eu



The example below shows the structure of the database entries:

- Domains: mycompany.eu
- Hosts: www.mycompany.eu, pics.mycompany.eu
- Directories: www.mycompany.eu/pics/
- HTML pages: www.mycompany.eu/pics/index.html
- Picture URLs: www.mycompany.eu/pics/001.jpg
- IP addresses: http://123.123.123.123
- Protocols: http:// vs. ftp://
- Ports: http://www.mycompany.eu:80 and http://www.mycompany.eu:81

If the URL is checked successfully, information is returned on which of the 58 categories it belongs to. The administrator can decide individually for the company which web sites or categories should result in what actions. The 58 categories are divided into 14 subject groups such as “Pornography, Nudity", "Shopping" or "Criminal Activities". You can activate or deactivate each of the categories that these groups contain. The sub-categories for "Criminal Activities" are, for example, "Illegal Activities", "Computer Criminality", "Political Extremism/Hate/Discrimination".

Content filter actions

The content-filter actions initiate the measures required for each user group, time and target address. Besides the usual actions of "allow" and "block", the LANCOM Content Filter provides a further option that can be used to handle access to the web sites in a flexible manner.

The "override" option enables the user to visit a web site that has actually been blocked. This option allows the administrator to avoid situations in which company interests make it vital to access certain web sites that would normally be blocked. In such cases there may be no time or opportunity to check access and to adapt the rules if necessary. If the user accepts the override, he or she may access the unblocked web

LANCOM Systems

LANCOM Techpaper
Content Filter

site for some minutes or even a whole day, depending on the configuration. The administrator can be automatically informed of such incidents in order, for example, to activate access for URLs that are subject to frequent overrides. The override type allows temporary unblocking to be defined more precisely. For example, the override feature can be used to unblock the category which the requested domain belongs to. This will then make all the other web pages belonging to this category accessible. Alternatively, the domain can be temporarily unblocked irrespective of the categories that the various sub-pages are assigned to. The combination of category and domain is especially restrictive for the override feature since only those URLs will be permitted that belong to both the domain as well as to the categories of the current web site.

The rating server

The rating server uses various web crawlers to evaluate the content of web sites. Web crawlers browse the Internet automatically and quickly, analyzing keywords, text and images in order to classify and store them in a database. The combination of various intelligent algorithms for text classification and image identification using techniques to identify, for example, symbols, logos, skins and faces as well as OCR (optical character recognition). This guarantees an extremely reliable rating. The database contains a list with over 100 million URLs that cover some 10 billion web pages and is updated and extended by the dynamic, automated web crawler process with almost 150,000 entries daily.

Web crawlers

The web crawler algorithms are specifically designed for the requirements of the rating server. At a technical level, the web crawlers must apply the right strategy to take account of performance fluctuations, unreachable servers, spam domains, parked domains, multi-lingual and search-engine optimized web sites. At the content level, cultural, moral and ethical aspects are important for the appropriate rating of web sites geographically. In order to be able to examine all relevant URLs, the web crawlers follow hyperlinks they find on the web sites – not within a domain but systematically beyond its boundaries into other domains. The web crawlers also make use of domain registration information and other external sources relating to new domains in order to find non-linked web pages.

Text classification

Text classification in the rating servers goes far beyond the simple examination of keywords. Even if it is easy to configure the search on the basis of keywords, a simple rating based on the number of occurrences of certain words has a significant disadvantage since some words have multiple meanings that can only be assessed under consideration of the context. The word "sex" can occur in, for example, pornographic as well as medical content.

For this reason, intelligent search functions use the frequency of words and the inter-combination of different words. These heuristic processes make it possible to classify texts extremely accurately, provided that the quantity of text is sufficient for a valid assessment.

Image classification

Several technologies are used to classify images on web pages. Pornographic pictures are determined, for example, on the basis of the proportion of skin tones in the images. The rating servers have a special facial recognition feature that can reliably identify faces and their proportion in the image in order to prevent portraits being incorrectly identified as pornography. But image classification can also recognize symbols with political or similar significance in order to identify prohibited content. As images can often contain text, optical character recognition (OCR) is used to classify any conspicuous messages.

LANCOM Techpaper
Content Filter

Combination of results

In some cases the results of the individual rating engines do not provide clear-cut findings. For example, a web page may display images with a large proportion of skin, which suggests pornographic pictures. However, the associated text indicates a medical context. The results are therefore summarized using a weighted rating that makes it possible to perform a reliable classification.

Automatic selection of the rating server

The LANCOM Content Filter uses a central rating server located in one of the world's largest computing centers. Besides the database, which is always up to date thanks to highly professional category management and the web sites examined, this solution offers an additional important advantage in that the LANCOM Content Filter will automatically use a backup server if there is any disruption to the rating server. There are several redundant databases available around the world, meaning that access to the current rating status is always available. The LANCOM Content Filter will always select the server with the fastest response time in order to optimize filter performance.

User-defined category profiles

The LANCOM Content Filter can be specifically tailored to an organization's requirements using user-defined category profiles. In many cases, blacklists and whitelists apply uniformly to all users – but the blocking or unblocking of different categories of web page does not need to be regulated uniformly for all parts of an organization. A press officer in the PR department might require access to URLs that should not be available to colleagues in production. In order to model these specific situations, profiles are set up that have their own category configuration. Such a user-defined profile is assigned to a user group via the firewall, which means that the user group then has its own filter settings.

Logging and alerts

The content filter's logging and alert functions are particularly important in light of the continuous development of filter rules. Depending on how it is configured, the content filter can inform the administrator about the number of licenses being exceeded or licenses expiring, about errors, or about blocks on web pages being overridden. The administrator can choose to be informed via e-mail, SNMP or SYSLOG. Furthermore, the content filter can create a snapshot at regular intervals to provide detailed statistics for a particular period for comparison with other periods. LANmonitor provides convenient functions that allow the status of the content filter to be viewed during operation.

Content filter statistics are always anonymous and do not allow any inference as to the user's identity. Only messages relating to the use of the override function can contain the IP and MAC address of the relevant computer and the requested URL, depending on the setting.

User-defined blocked pages

The LANCOM Content Filter provides predefined pages for the "blocking" and "override" functions that can be displayed in the browser when the actions are triggered. A company can develop its own blocking and override pages to allow the content filter to be fully integrated into the organization. All modern web-page functions such as JavaScript are available for this. Simple HTML tags, which are inserted into the source text at run time, can be used to display function-related texts and buttons.

LANCOM Techpaper
Content Filter

LANCOM content filters are used to implement corporate-wide policies

Administrators can use a content filter to systematically filter content in the network and thereby prevent access to Internet sites with content that is illegal, offensive or hazardous. This can also be used to prevent private surfing during working hours. LANCOM content filters allow Internet access to be centrally configured and managed for the entire organization, allowing Internet policies to be implemented in a standardized and effective manner. This increases employee productivity and makes the network more secure, reduces the risk of illegal activity in many areas and reserves the entire bandwidth for business processes only.

The LANCOM Content Filter allows the network to be tailored to match organizational-specific requirements using a whole range of functions including the specific restriction by group, IP address or workstation, scheduled rules and warnings, or the complete blocking of certain web pages. The basis for filtering is provided by a central database, which currently contains more than 100 million URLs. These URLs are grouped into 58 categories that assist with customer-specific configuration.

The continuous maintenance of the database goes far beyond the concepts of many other types of solution. The effort of manually maintaining URL lists can only ever be partially successful because the resulting errors can lead to gaps in network security. The LANCOM Content Filter uses a database that automatically examines web sites, meaning that new and modified web content can be integrated very quickly. Web crawlers update and extend the central database with almost 150,000 web sites each day using complex analytical methods for automatic categorization. Using the LANCOM Content Filter, the administrator can easily make informed decisions based on categories about what Internet content is allowed in the network and what is not.

Summary

For many companies access to the Internet is a business-critical process. Implementing clear and uniform policies for Internet usage is essential in order to guard against the dangers of malware etc., to ensure employee productivity, and to meet legal requirements. The LANCOM Content Filter makes it possible set up a central, powerful filter solution for corporate networks that is also easy to configure. Using a database that is always up to date and rules specifically tailored to individual requirements, the LANCOM Content Filter allows the effective implementation of company-wide policies, thus increasing the security of the network.