Multilayer Dynamic Internet Content Filtering

An Overview of Next Generation Filtering Technology

Table of Contents

1. Introduction

A growing number of businesses, organizations and households are actively managing the Internet content that their employees, members, and children can access on a daily basis. The deployment of Internet content filtering solutions is on the rise. Even in today's tough economy, organizations recognize the potential for filtering to save significant costs and redeem lost productivity. Educational institutions must fulfill their responsibility to provide safe Internet access for students; and parents want to make sure their children are not exposed to inappropriate content at home.

Managing Internet access is a tremendous challenge. Internet connectivity is prevalent in the work place, at school and in the home. An abundant variety of content is available today, and the Internet continues to grow exponentially. No matter what the inappropriate content may be, businesses, schools, and parents want to be sure they can manage access to it via their networks or Internet connection.

The dynamic nature of the Internet requires a filtering tool that can match its pace of growth and change. In the long run, and even today, only a multilayer dynamic content filtering solution can truly "keep up" with the constantly changing make-up of the Internet and the World Wide Web. It is essential for organizations and individuals who are serious about filtering to match the tool they use to the formidable task at hand.

2. Internet Then and Now

In the earliest days of the Web, there were relatively few documents and sites. It was a manageable task to post all documents as static pages. Because all pages were persistent and constantly available, they could be found and easily categorized by conventional search engines and web crawlers. In July 1994, the Lycos search engine went public with a catalog of 54,000 documents. Since then, there has been exponential growth rate in available Web documents making it virtually impossible to maintain accurate figures defining the actual quantity of information available!

Internet content is considerably more diverse and the volume much larger than commonly understood. Most Internet surfers are aware only of the content presented to them via search engines such as Excite, Google, and AltaVista, or search directories such as Yahoo! and About.com. To be discovered, the page must be static, linked to other pages, and accessible for 1-2 months before it will be indexed by these search engines. When many analysts and researchers do discuss the size of the WWW, they are commonly referring only to documents in what is known as the "Surface Web." To really understand what you are facing, in terms of the size and the continued fast paced growth of the WWW, it is important to recognize that there is a "surface web", which comprises content that can be accessed by the search engines, and there is a "deep web" whose size greatly exceeds the "surface web."

2.1. Surface Web

The surface Web contains an estimated 8 billion documents, growing at a rate of 7.5 million documents per day! Surface Web sites usually contain a number of fixed HTML pages posted within a static directory structure that comprises a home page with links to sub-pages. Many commentators have noted the increasing delays in posting and recording new information on conventional search engines. Empirical

tests by search engine vendors suggest that listings are frequently three or four months – or more –out of date. And we've only scratched the "surface." In truth, most of the Web's information is buried deep in dynamically generated sites based on database-driven designs—and standard search engines (i.e., crawlers) never find it.

2.2. Deep Web

The deep Web is qualitatively different from the surface Web. Deep Web sources store their content in searchable databases that produce results dynamically in response to a direct request. In other words, deep Web pages do not exist until they are created on the fly as the result of a specific query. Users who know the information is there, will know how to get it. Users who are searching for information would have to make direct queries—one at a time—until they hit upon the right request. This is like finding a needle in a haystack, assuming you know that the haystack exists!

Public information on the deep Web is currently estimated to be 400 to 550 times larger than the commonly defined World Wide Web. The deep Web contains 7,500 terabytes of information compared to 19 terabytes of information in the surface Web. It is now accepted practice for large data producers and new classes of Internet-based companies to choose the Web as their preferred medium for commerce and information transfer. However, the means by which these entities provide their information is no longer through static pages but through database-driven designs.

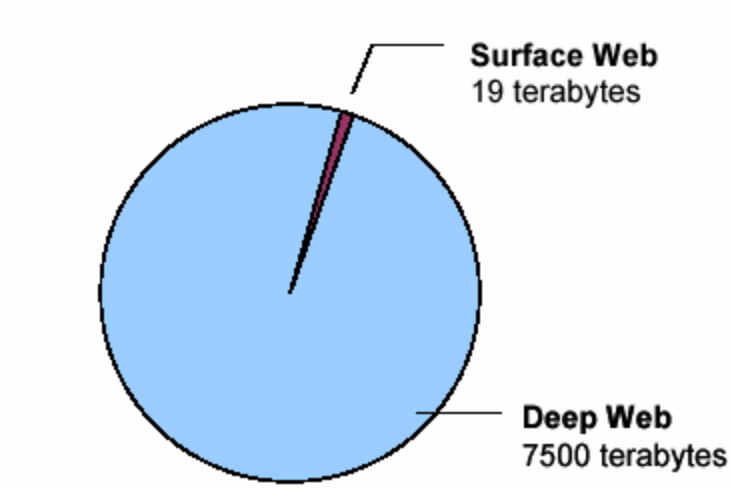

Figure 1: Terabytes of information in surface Web and deep Web

Clearly, the deep Web is the largest growing category of new information on the Internet. On average, deep Web sites receive 50 percent greater monthly traffic than surface sites and are more highly linked to than surface sites. Deep Web sites tend to be narrower, have deeper content and are highly relevant to every information need, market, and domain. A full 95 percent of the Deep Web is publicly accessible information—not subject to fees or subscriptions.

2.3. How many URLs are there?

The estimated number of URLs on the Web fluctuates widely because no one really knows how many there are. Current wisdom says whatever the current number, it's much more than you think. Some educated estimates* are presented here.

• July 2001 survey data from the Internet Software Consortium provides estimates that there were well over 125 million Web sites at the time of the survey

• 4,285,199,774 web pages were found in a March, 2004 Google web search

• In November 2004 Google's index has reached nearly 8 billion pages

• By conservative estimates, the deep Web contains 400 times more information than the surface Web. If we compare terabytes to terabytes, the "surface" has 19 while the "deep" holds over 7500

• In July 2001, the Internet Software Consortium counted 125,888,197 Internet domain hosts. By January 2002, that figure had grown to 147,344,723 — a growth of 17% in only six months. By January 2004, Internet domain hosts figures had reached 233,101,481

• Estimates place the percentage of adult sites on the Internet as high as 25% of available content

• Daily pornographic search engine requests have reached 68 million in 2004 (25% of total search engine requests)

• In 2004, Adult sites have generated an estimated $2.5 billion in revenues world-wide

Quantifying the number of URLs loses its significance when we consider a study conducted at the NEC Research Institute and published in *Nature* in July 1999. The study estimated that the search engines with the largest number of Web pages indexed (such as Google or Northern Light) each index contained no more than sixteen per cent of the surface Web. Since they are missing the deep Web, these search engines (and their crawlers) are therefore searching only 0.03%—or one in 3,000—of the pages available to them today.

3. Managing Access is a Must

3.1. Managing Internet Access in the Enterprise

Corporations are increasingly aware of the advantages and protection that filtering offers. Both large and small businesses stand to lose employee productivity and suffer financial damages from lawsuits. In addition, they bear the rising cost of bandwidth needed to support unbridled Internet access. They seek a comprehensive content filtering solution to help them manage their business needs. Some researchers estimate that as much as 80 percent of Internet use during business hours is not work-related, turning this important business tool into a productivity hindrance.

When employees access the Internet for non-work related reasons, they needlessly occupy corporate bandwidth with bandwidth-hungry content such as video streaming, heavy graphics, and music files. Surveys have shown that at-work surfers are heavier consumers of on-line media than at-home surfers.

Managed use of bandwidth can reduce the IT administrative burden and the amount of IT equipment needed. It also lowers the overall cost of Internet access, which can be quite significant, especially in large networks that support many users who are probably misusing their Internet access on a regular basis. Furthermore, Internet filters are able to prevent the downloading of questionable content, and therefore, limit corporate liability for sexual harassment lawsuits.

Many companies have felt the effects of sexual harassment litigation, which can cost millions of dollars in settlements, legal fees, and a tarnished corporate image. The mere existence of certain types of content on the network (such as pedophilia) is a criminal offense in some countries. Recent examples of this were noted by the BBC when they reported that UK telecom service provider, Orange, fired 40 workers for electronically distributing pornographic material over the company's network (BBC, July 9, 2002). Also, in August 2004, following an internal investigation at the UK DWP (Department for Work and Pensions) one person has been prosecuted and 277 people disciplined, including 16 who were sacked, for inappropriate computer usage. A survey by UK based Personnel Magazine found that Internet abuse tops the list of causes for disciplinary action in UK companies. (Sept 2 2002, Reuters: *Web abuse main reason for disciplinary action*). Likewise, a 2003 study published by IT Week found that half of UK firms have taken disciplinary action against employees because of Internet misuse in 2002,

Even though network managers may be unaware that an employee is downloading illegal content onto the organization's network, the organization is still held responsible for hosting the illegal content. By taking action to prevent all such content from entering the network, the organization fully protects itself from legal action.

3.2. Managing Internet Access in Schools

Educational markets are motivated to manage Internet access due to an increasing awareness of the dangers faced by minors engaging in unmonitored Internet usage. The main drivers for content filtering in educational institutions deal mainly with protecting children from inappropriate material and people.

According to a 2003 study by the European Opinion Research Group, 31% of European children access the Internet at school. This generates an overwhelming concern from parents and teachers regarding children viewing pornography, hate, drug sites, and other inappropriate material. This issue has fueled the market for content filtering products in educational and residential settings. Just as a school library would never intentionally provide inappropriate books or magazines to its students, it is equally improper for students to have access to inappropriate content via the school's Internet connection.

Furthermore, parents and teachers are increasingly concerned about the dangers children face when they naively visit seemingly innocuous Web sites that actually contain adult content. It is common practice for owners of illicit content to "kidnap" URLs when their domain registration lapses, or to use common misspellings of popular sites to avoid exposing the true nature of their content, but at the same time, to ensure numerous, albeit unintended hits. For example, in October 2002, we identified the www.historyplaces.com site as a pornography site. Another site,

www.historyplace.com is a legitimate history education site. Obviously, the owners of the pornography site purposely chose a deceitful name. There are many examples of popular sites that allowed their registration to lapse just momentarily, and were taken over by pornography sites. Perhaps one of the best known of these is www.whitehouse.com, which is a frequently visited pornography site. (http://oii.org/html/porn-napping.html).

US Federal legislation, such as the Child Information Protection Act (CIPA) will drive content filtering in educational institutions across the United States. The CIPA requires content filtering in schools and libraries that are supported by US federal funding sources. Similar laws and regulations have been enacted in Europe, Australia and other regions.

Internet Pornography Statistics show that the Average age of first Internet exposure to pornography is 11 years old. Figures show that 90% of children aged between 8 and 16 have viewed porn online. Likewise, studies have found that 80% of teens aged 15 to 17 have had multiple hard-core exposures.****

3.3. Managing Internet Use at Home

The benefits created by the development of the Internet are without question. However, the Internet has also become a medium to communicate things that many households would consider dangerous, offensive, sexist, racist, or otherwise inappropriate. A content filtering solution is one way for home users to take the initiative to create their own safe Internet environment.

Internet Service Providers (ISP) have entered the filtering arena in full force as subscribers lean more and more toward holding their access provider responsible for the content available to them. By offering filtering as a value-added service, ISPs give parents the ability to manage the type of content their children can access. Moreover, places as different as South Dakota and Australia are requiring Internet Service Providers to provide filtering services and to report hosts who are peddling prohibited content over their networks.

4. Filtering Methodologies

There are two principal methods for identifying and filtering Web content as users browse the Internet: URL Databases and Dynamic Content Analysis.

4.1. Static URL Filtering: URL Databases

Proprietary collections of URLs are designed to associate each URL they contain with a specific content category that users may want to filter. When a site is requested, the filter looks up the address of the requested Web site in the database. By comparing the requested URL against the database, the filter can block or allow access to the site in accordance with the Internet usage policy set up by the organization or individual.

URL databases are supplied and maintained by the filter vendor. The database contains entries for Internet domain names and specific sub-domain URLs. Each entry is classified into a specific category of activity. URLs not found in the database are usually allowed through, although most filters can be configured to block all uncategorized traffic.

The conventional wisdom about filters that use a URL database is that they offer good scalability but limited coverage. Typical databases provided by the leading vendors do not include more than five million URLs. Vendor-supplied databases are neither complete nor completely accurate, but they do provide a way to quickly categorize sites for blocking or reporting purposes. They do a good job at collecting and categorizing established and frequently visited sites. They also have historically performed well from a scalability standpoint because filtering is a simple matter of a URL lookup against a database. However, database filters are beginning to feel the performance crunch because the larger they get, the longer it takes to conduct a lookup and to update. Eventually, databases may not be able to expand further without severely affecting performance.

Building a URL Database

URL databases are supplied and maintained by the filter vendor and are updated regularly, usually on a subscription basis. Each entry is classified into a specific content category such as drugs, gambling, hate, pornography, and many others. Classification is done by researchers who manually review and categorize each URL. Customers can usually augment the database with their own entries or new categories.

A variety of methods are used to build the database. Clearly, there is no precise methodology for mining the Internet and some vendors have developed better methodologies than others. Since no database will realistically include all of the sites that the employees in an organization could potentially access, some vendors include an option for customers to send all uncategorized sites to the vendor for review. This way, over time, the database will more accurately reflect the surfing habits of the organization. However, research indicates that most organizations are highly reluctant to open their networks to such a security risk. This method reveals every site being accessed from within the organization. It is proprietary data most companies are not willing to release.

The vendor's researchers also use search engines to find sites in certain categories. As previously noted, there are severe limitations to this method. New sites are available long before they make it to the search engine indices, and most pages on the Internet today are part of the "deep web" and never get indexed by the search engines.

Maintaining a URL Database

Once sites are mined, a human review process assigns each site to a filtering category. This is a time-consuming process. Maintaining a fresh database is a challenge that some vendors have addressed by developing automated tracking tools. Periodically, the content of the URLs in the database must be compared to the content currently available online. Automated tools help the vendor keep the existing list accurate so when content changes, it can be re-reviewed and re-categorized if necessary. Obsolete sites will be removed.

One leading URL database solution provider explains that they have 40 human researchers with a range of language skills responsible for adding 35,000 sites per week to their database. They also manually verify that the content of existing sites in their database has not changed. The methodology described is vague and despite the impressive statistics provided, they continue to rapidly fall behind the growth of the Internet:

A study by Cyveillance revealed that in the year 2000, there were 2.1 billion unique HTTP URLs on the Internet with an average of 7.3 million added daily. The study was

made over an eight-month period to determine the rate of Internet growth. The conclusion was that Internet growth had not reached its peak and the growth rate was actually accelerating.*

4.2. Dynamic Filtering: Dynamic Content Analysis

Dynamic content inspection performs on-the-fly content analysis of the Web traffic as it enters the internal network. Different techniques for content analysis include context-sensitive text analysis, inference engines, and neural networks. When the Web page is received, it is analyzed and then categorized according to the content found in the page. After the content has been scanned, the system determines whether to block the transmission or simply log the activity under a particular category.

In the past, critics have contended that while Dynamic Content Inspection provides good coverage of the Internet, it has inherent latency that can degrade browsing response time. But today, there are content analysis engines that add negligible latency and have even less effect upon network performance than URL database solutions. This enables customers to benefit from greater coverage of the Internet without network performance degradation.

A dynamic filter becomes accurate by defining a category very precisely and training the engine to accurately identify the sites that fit the category. Using many parameters, the dynamic filter will be able to distinguish, for example, between content that advocates drug use and content that teaches about the problems associated with drug use. Within a site, it can distinguish individual pages that are appropriate and others that are inappropriate instead of blocking the entire site as a database filter would do.

The main difference between the two methodologies lies not only in technique, but also in relevance. URL databases worked well when the Web was smaller and contained a measurable number of static sites with linked pages. But today, the URL database method is woefully inadequate in its ability to cover the proliferation of Surface and Deep Web sites. Given the inability of any vendor to actually maintain reasonable pace with the rapid growth of the Web, it would seem that the URL database is a filtering method whose time has come and gone. It is interesting to note that in their 2001 report, *Content Filtering*, Frost & Sullivan predict that solutions based on artificial intelligence will eventually be introduced into the market and cause a surge in the sector's growth when enterprises move to upgrade to the new generation of filters.**

5. Attributes of an effective Content Filtering solution – The Multilayer approach

In the following paragraphs, we will show how a combination of static filtering and dynamic inspection is the only filtering method that provides real accuracy and scalability as the Web weaves an increasingly sophisticated network of sites.

5.1. High Accuracy

The quality of any Internet filter must be judged on its ability to accurately determine the type or category of content on a given Web site. Inaccurate filtering produces two very undesirable outcomes: *false positives* and *false negatives*. False positives are produced when the filter blocks access to valid content because it incorrectly identified it as impermissible. False negatives are produced when the filter

allows access to inappropriate content because, once again, it failed to categorize it correctly. Both types of false results have negative impact on an organization's content management policy. False negatives impede productivity and frustrate users, while false positives can expose the organization to legal liability. Likewise, false results mean the filter is failing to accurately enforce the organization's Internet usage policy.

Dynamic vs. Database Accuracy

Theoretically, a URL database solution always correctly filters the URLs in its database. The caveat here is that the filtering is only as good as the collection of URLs in the database. As we have discussed earlier in this paper, even the biggest URL database covers only a fraction of the Surface Web and virtually none of the Deep Web. Since these proprietary URL lists and databases are compiled by a combination of Web crawlers and human reviewers, it is quite impossible to find, review, and categorize enough of the available Web sites to keep the database current.

In contrast, dynamic filters analyze and categorize Web content on the fly. Whether a Web site has existed for 5 months or for 5 minutes is not important, because the determination of the category in which the Web site belongs is made just at the time of the request. Therefore, dynamic filters have no problem keeping up with the growth and changing content of the Internet.

The often-heard criticism regarding dynamic filters is that they are not as accurate as they need to be and therefore, they generate too many false positives and false negatives. This criticism may have been true of early commercial dynamic filters that were limited to analyzing keywords. But the sophisticated implementation of newer technologies has produced dynamic filters that are far more accurate from the get-go – and can be trained to perfect their classification capabilities over time.

Dynamic filters, like PureSight are not subject to human error, which happens frequently when reviewers have to scan and read millions of Web pages every day to keep their URL database current. For example, some pornographic web sites hide their content behind an innocuous home page and even second page. If a human reviewer doesn't go deep enough into the site, it could be classified incorrectly.

The PureSight dynamic filter takes into consideration hundreds of parameters when analyzing a web page and it examines the co-dependencies of those parameters, including everything from graphics, to fonts to colors used on the page.

5.2. Up to date

A content filter must be up to date with what is available on the Internet in order to be effective. Multilayer dynamic content filters are always up to date. There is no lag time between finding and categorizing sites because each page can be analyzed anew at runtime. In contrast, URL databases constantly struggle to stay updated. They employ many aids in this task, including Web crawlers to find new sites, and teams of researchers to review and categorize the sites.

Web crawlers

A crawler gathers web documents and indexes them. As the web grows, so does the amount of data that needs to be "crawled" through and the crawler must scale to

keep up. In order to scale to hundreds of millions of web pages, many fast distributed crawling systems have been developed in which a single URL server serves lists of URLs to multiple crawlers, each keeping roughly 300 connections open at once. This is necessary to retrieve web pages at a fast enough pace.

Even though crawlers are improving constantly, the technology they employ is no match for the sheer speed and diversity of Web growth. They hop from site to site and from one URL to another, indexing the contents of pages as they go. It's slow going. And when they bump into sites where information is held inside a database, they grind to a halt. Moreover, once a page is indexed by the crawler, it still must be categorized in the URL database and this process could take weeks.

Customer Updates

AI (Artificial Intelligence) engines allow a greater independence as dynamic filters are never down due to update procedures. In contrast, URL database solutions require the user to receive updates on a regular basis in order to keep the list as current as possible. This can be as often as once a day. The update procedure downloads all the new and updated classified URLs from the vendor's master database to the user's network. The amount of data to be downloaded can be significant, causing heavy load on the network. Although updates are normally scheduled for off-hours, today's businesses have employees working all over the world, accessing the network around the clock.

In contrast, a multilayer dynamic filter such as PureSight is not concerned if a URL has existed for only one minute – the dynamic filter will inspect, categorize, and allow or deny access on the spot according to the user's predefined access policy.

5.3. Preventing Work-Arounds

As filtering and monitoring Internet activity becomes more prevalent in organizations, workaround solutions have become available and surfers have found ways to circumvent some of the filtering solutions deployed on their networks.

• **Anonymizers** are applications and proxy services that prevent filtering software from determining which URLs are being accessed. This makes a database filtering solution completely ineffective.

• **Content Distribution Network (CDN)** service providers like Akamai Technologies use HTTP redirection techniques, which modify the source URL of the content. In these cases, a database filtering system will not work unless the "altered" URLs that are associated with the CDN service are also blocked.

A multilayer dynamic filter that analyzes content on the fly is not affected by changes in a URL and cannot be bypassed by methods that alter the requested URL address. Most dynamic filters and certainly the PureSight multilayer dynamic filter analyzes *every* Web page (packet) entering the network regardless of its URL address.

Not Classified

When a requested Web page is not in the URL database, the page is "unclassified" and the organization must decide whether the user will be permitted to view this page or will be denied access.

• If the organization allows access to unclassified sites, it will be under-blocking since it is inevitable that many unclassified sites will fall into categories that the organization's Internet Usage Policy defined as impermissible.

• If the organization denies access to uncategorized sites, it will be over-blocking since literally millions of sites are not in the URL databases and it is reasonable to assume that some users will need to gain access to some of those uncategorized sites.
In contrast, a multilayer dynamic filter such as PureSight addresses every site that is accessed.

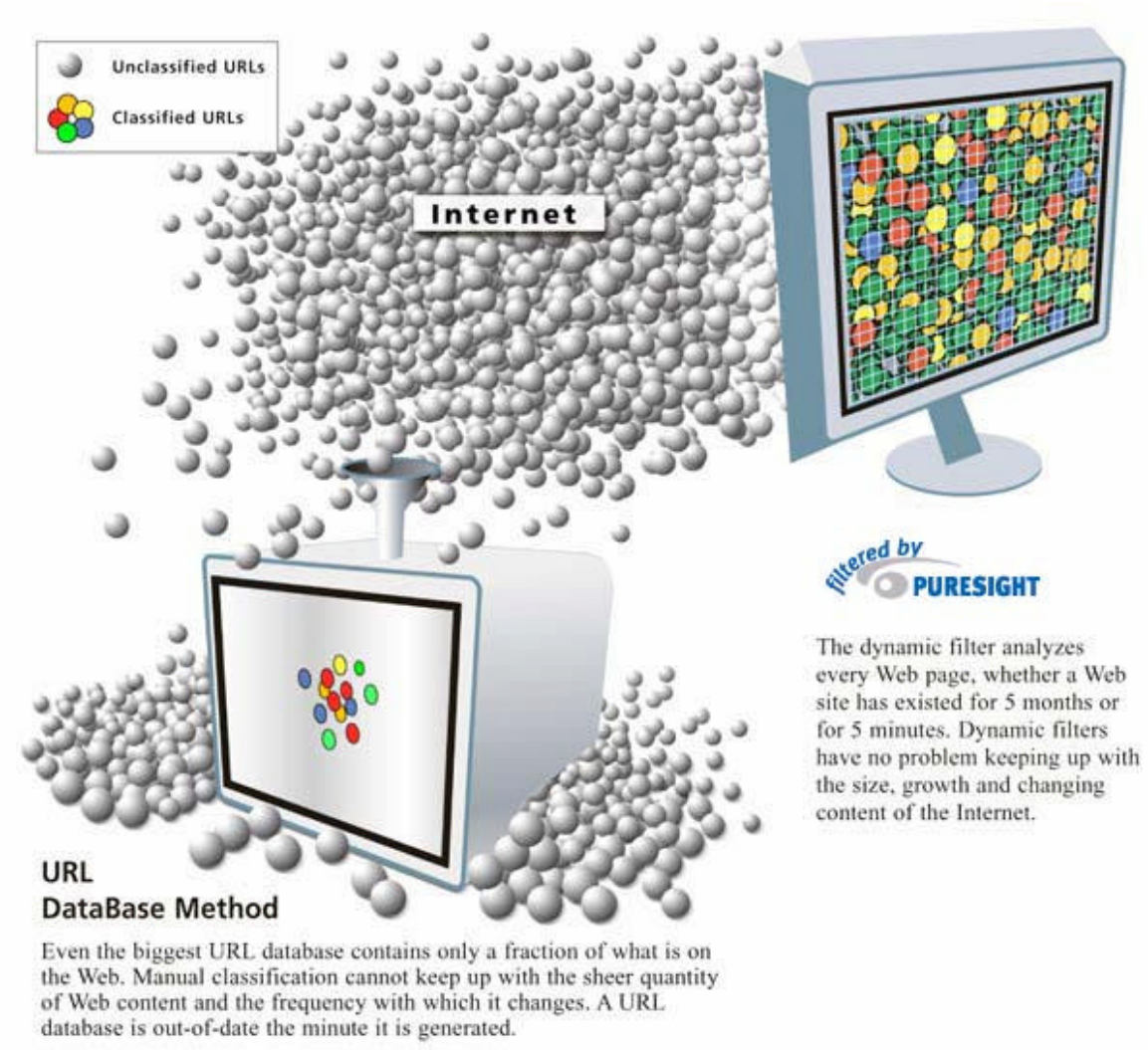

Figure 2: Multilayer Dynamic filters cover every URL on the Internet, while Static Database filters cover only a small fraction of URLs at any time

Filtering Stolen and Innocently Named URLs

Domain names that have lapsed or are no longer in use are subject to "theft" by content providers who are happy to hide behind the innocuous name precisely because it is unrelated to their line of business. Formerly legitimate web sites such as http://www.anna.com (American Nephrology Nurses Association) went from informational to adult content, without changing the URL.

Likewise, innocent URL names often lead to objectionable information the user was not expecting. This happens quite often to youngsters who search on common words like “kitten” or “girls” or they slightly misspell the name of a popular rock star and end up at a sex site. For example, www.britneyspears.com takes you to the site of the well-known music star by the same name. However, www.britnyspears.com is a site that contains adult material.

A dynamic solution will handle these sites no differently than any other web page containing content that is not in line with the organization’s Internet usage policy. The filter cannot be “fooled”. The problem a database solution might have is in maintaining pace with the changes (i.e., URL thefts). These sites may eventually be added to the database, but the question is, when?

5.4. Multilingual Support

The Internet is a global community where theoretically, access to information is unlimited, no matter where it originates or in which language it is written. Most database filtering solutions are highly English-centric. It is difficult to analyze how much coverage of non-English web sites is provided by the popular database filtering solutions because the vendors do not publish their lists and do not provide statistics concerning the number of sites in a particular language. Therefore, claims about multi-lingual coverage are vague since having a few sites in Russian, for example, may be all that is supporting the claim that Russian language sites are covered. The main challenge to providers of database filters is the need to hire a multi-lingual staff to accurately review sites in each language. Of course, this can be highly cost prohibitive.

A dynamic content filter, like PureSight, is multilingual by nature because its engine can be trained to inspect and filter any language. There is no need to employ language capable staff. The engine, when trained correctly, will learn to discern text and interdependencies between specific words and phrases, making it able to identify the content accurately, no matter what language is used.

5.5. Filtering Password-Protected Content

Access to password-protected sites is available only by payment or some other membership qualification. URL database solutions are not capable of registering or paying membership fees for the multitude of protected/private sites that are available and fee-based content is particularly popular among pornographic sites. Furthermore, it is often difficult to determine that a site contains adult content without getting beyond the registration page. In these cases, the dynamic filter offers a distinct advantage because regardless of whether the site was accessed using a password or not, the content is analyzed and categorized as it streams into the user’s network.

6. Multilayer Filtering – the PureSight Way

PureSight uses several filtering layers and technologies, including automatic and manual categorizations.

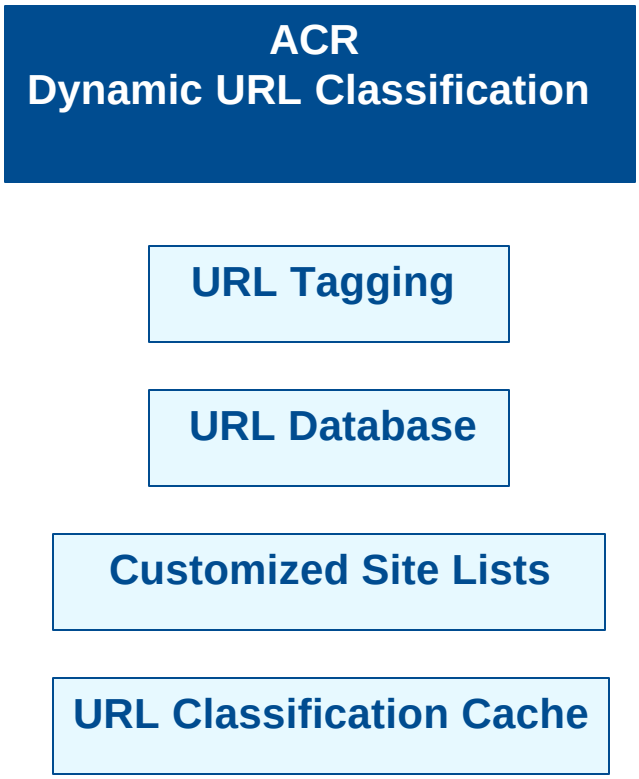

Figure 1: Multilayer Dynamic Content Filtering

The PureSight Multilayer Approach

PureSight offers an innovative multilayer filtering approach by combining the PureSight Artificial Content Recognition (ACR) core technology to a pre-classified database, customized site lists, URL cache categorization and tagging capabilities.

The ACR examines every page request to ensure its compliance with corporate, institutional, or parental acceptable use policies. The ACR powerful set of algorithms analyzes and categorizes data in real-time. The database, the static URL lists, the URL tagging and URL cache categorization are complementary filtering layers guaranteeing total coverage of the Internet.

The PureSight filtering layers:

- ACR™ – "On the Fly filtering" - dynamic filtering of all requested pages based on advanced content recognition technology
- Database – black list of sites classified by category remotely managed by PureSight Inc.
- Network static list – local, customized black/white lists of forbidden/allowed lists managed locally by the network administrator related to specific Internet access policy set for the users
- Cache categorization – local cache of previously classified URLs for network performance enhancement
- Tagging – identifies content category via the standard content rating systems used today (PICs, iCRA, SafeSurf)

6.1. One-pass intelligent ACR system

The PureSight dynamic filter core technology is a sophisticated proprietary Artificial Content Recognition (ACR) technology, that can "identify" the content of a site and then decide whether to allow it or not. Every requested site is inspected by PureSight's intelligent algorithm to ensure its compliance with corporate, institutional, or parental usage policies. PureSight provides complete and reliable Web coverage with unmatched recognition accuracy.

ACR comprises a powerful set of Artificial Intelligence algorithms that analyze and categorize data in real-time. Here's how it works.

Parsing (from Web page to parameters to Raw Data Vector)

Any web page requested by a user is received packet by packet and sent to the HTML Parser. The parser is the first component of the ACR that inspects the incoming data. When packets arrive, the parser breaks down the HTML code into hundreds of parameters, creating a long vector called a Raw Data Vector (RDV).

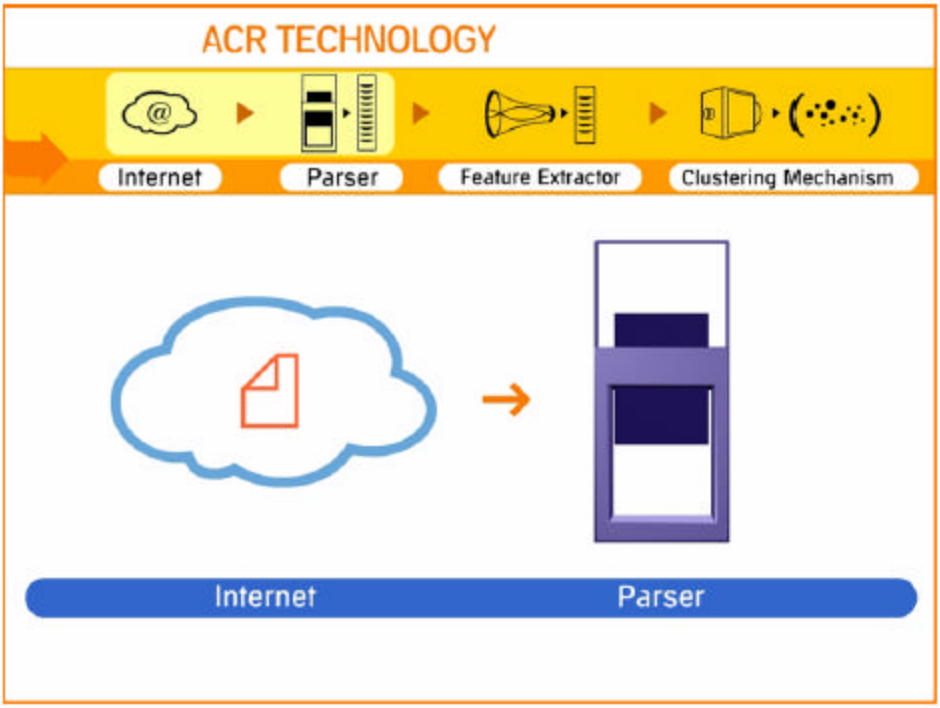

Figure 2: Requested Web page is sent packet by packet to the Parser

Parameters used in Parsing

The ACR engine recognizes different types of parameters and takes into account the words used in the page and the basic layout and format of the page. For example, pornography pages typically have very similar characteristics, explicit words, a large number of graphics, dark backgrounds, large fonts, etc. In contrast, educational sites containing information about sex education typically contain extensive texts, formal language, light backgrounds, simple fonts and few graphics. These differences in appearance and language are what enable the ACR engine to distinguish between the two very different categories of content.

Following is a partial list of the parameters analyzed by the ACR engine:

Non-textual information:

- Background colors
- Type of font (used in text, in headers)
- Color of fonts
- Font size
- Number of links
- Number of pictures
- Size of pictures
- Number of frames
- Average size of words
- Number of words

• Special signs

• Meta tags

Textual information:

• URL name

• Meta tag text

• Dictionary Words

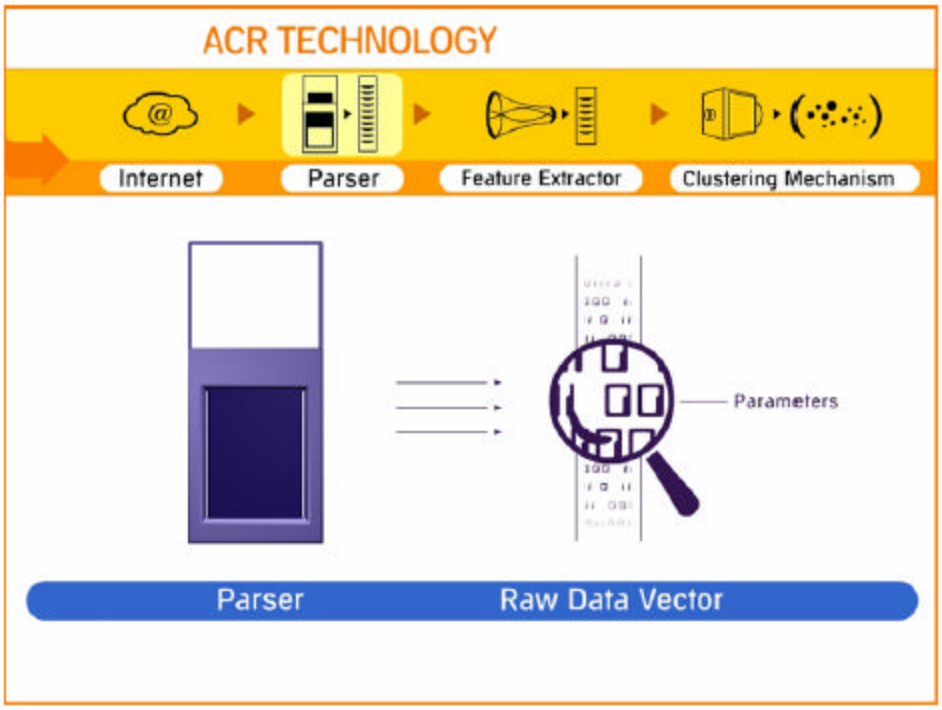

Figure 3: Parser breaks the Web page into hundreds of parameters in a Raw Data Vector

A list of about 20-30 dictionary words are selected and encoded into the Parser for each category. PureSight experts select and encode dictionary words for each category. The total number of dictionary words occurring in the site is one parameter among many in the RDV. Likewise, the HTML context of a dictionary word is also noted. Did the word occur in a meta tag, in a link, or in plain viewable text? More than the word itself, these relationships contribute to accurate analysis of the content.

For example, the Raw Data Vector might contain as one parameter, a list of words corresponding to a dictionary of sex terms; as a second parameter, a list of words corresponding to a dictionary of sports terms; while a third parameter would be the total number of dictionary words found on the HTML page. However, the existence of

a parameter for the sex words does not influence the Feature Extractor and the Clustering Mechanism to classify the page in a sex category. Rather, the PureSight program analyzes the relationship between the sex-terms parameter and the sports-terms parameter, and other parameters in order to differentiate between web page with adult content and a web page with sports content.

Extracting Features from Raw Data Vector to create a Processed Data Vector

The Raw Data Vector is processed by the PureSight Feature Extractor, which is the first layer of artificial intelligent algorithms. The Feature Extractor reduces the RDV from hundreds of parameters and creates another vector containing 20-25 specific features, called the Processed Data Vector (PDV). Specifically, the Feature Extractor finds patterns and inter-dependencies among the RDV parameters that are useful in classifying the web page.

For example, the Feature Extractor might compare the color of the words to the color of the background to obtain one such pattern. In this way, the parameters of the Raw Data Vector are reduced to tens of features that can be efficiently analyzed. In other words, the Feature Extractor associates one RDV parameter with other RDV parameters according to its AI rules and extracts relationships that are used to identify the content.

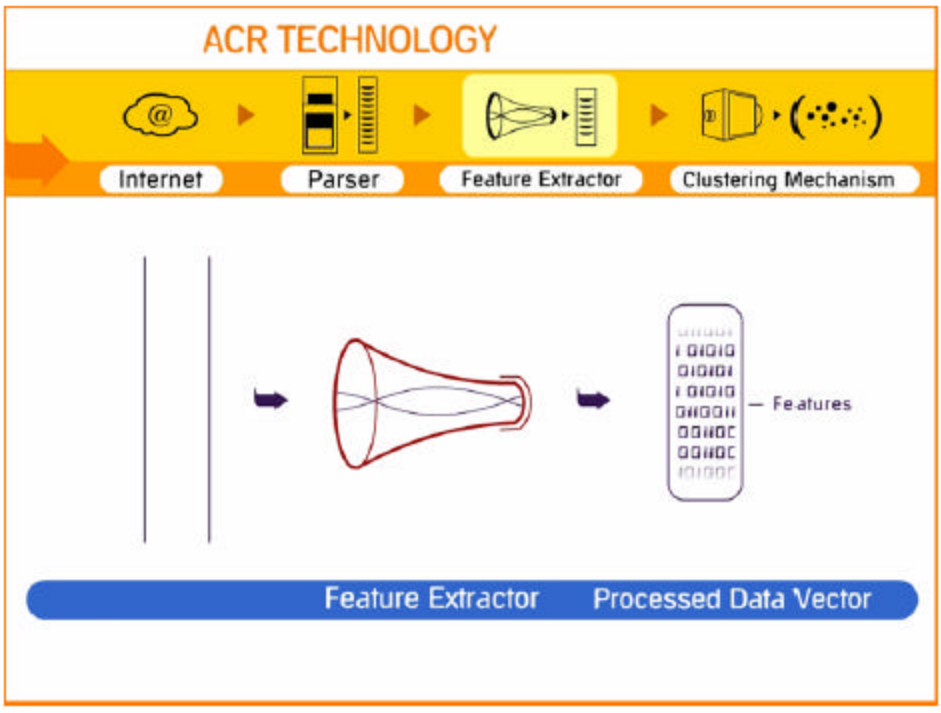

Figure 4: The Feature Extractor reduces hundreds of parameters into tens of features within a Processed Data Vector

Clustering – Assigning the PDV to a Content Category

Next, the PDV is processed by a second layer of AI algorithms called the Clustering mechanism. The Clustering mechanism is a neural network that analyzes the combinations and relationships in the Processed Data Vector, and extracts a mathematical coordinate for the HTML page. The mathematical coordinate is a coordinate in multi-dimensional space, which is "occupied" by several "clouds' of coordinates. Each cloud corresponds to a particular category of content and the Clustering mechanism knows to which cloud (i.e., category) a page belongs according to the value of its mathematical coordinate.

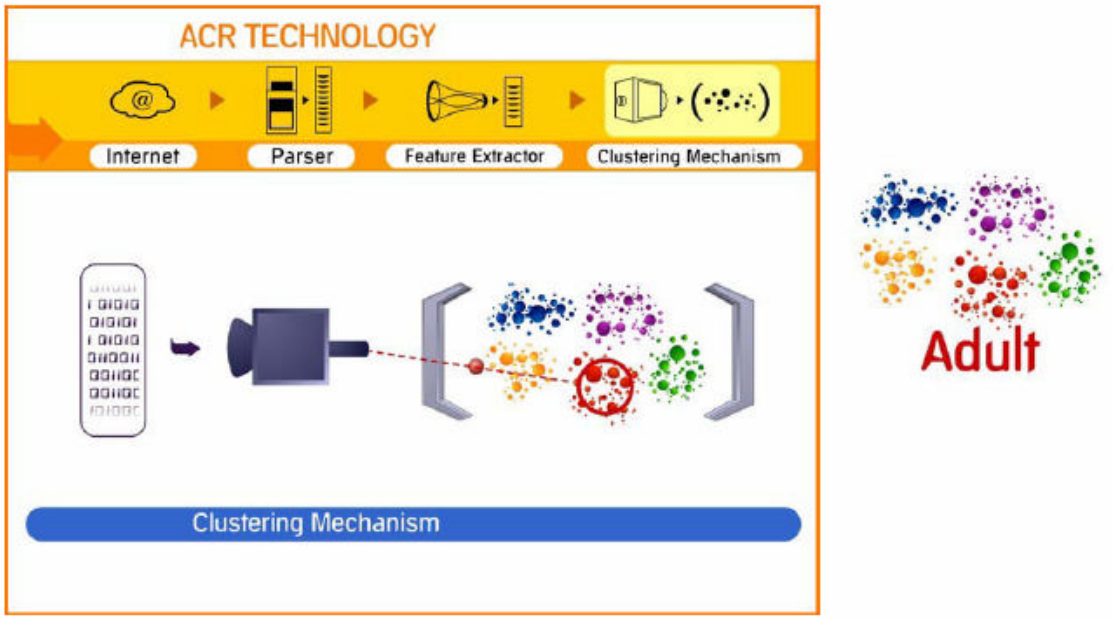

Figure 5: Clustering Mechanism analyzes and processes the PDV into a mathematical coordinate that is located in some proximity to a category cluster

For instance, one cloud of coordinates represents adult entertainment, while another cloud represents sports, and yet another cloud represents gambling. By matching the mathematical coordinate to one of the clouds, the ACR filter engine is able to identify the type of web page being requested by the user.

The ACR engine categorizes the webpage if there is a greater than 80% correlation with a cloud for one of the pre-classified categories of content. If the cloud represents a content category that is deemed inappropriate by the Internet usage policy, the webpage is not displayed. Otherwise, the webpage is classified as acceptable content and is displayed.

6.2. Making ACR Accurate

PureSight's ACR engine undergoes rigorous training in PureSight's quality assurance laboratories to ensure accurate filtering.

Training the ACR engine

Training the ACR engine is best described by the following example. To train the ACR engine to identify and distinguish between pornography and sex education content, our testing engineers feed the engine with sample HTML pages from both categories. The engine is "told" which sites belong to the category being defined (in this case pornography/adult) and which sites do not fit the category (in this case, sex education sites). Once the engine learns this "lesson" it can then make the fine distinctions, on the fly, between content that may have similarities but essentially belongs to very different categories.

In addition the ACR engine can be trained to identify content delivered in different languages. PureSight engineers encode a language-specific set of dictionary words for each category into the parser, and then define the appropriate interdependencies in the Feature Extractor.

6.3. ACR Performance

Latency

The ACR engine utilizes a one-pass inspection technology that analyzes packets upon arrival rather than waiting for the entire page to be downloaded, so processing is extremely fast and with no noticeable latency to the end user. In addition, the PureSight solution uses a URL Cache to store classifications of previously requested sites that were analyzed by the ACR engine. The URL Cache increases performance and reduces latency by actually building up a dynamic database that accurately reflects the surfing habits of the organization. The database therefore does not include sites that are never accessed by the network users and therefore are irrelevant for the organization. This is very different than the URL database solution, which will by definition contain many URLs that are never accessed – and are therefore not useful.

General Classification versus Specific Categories

PureSight has designed PureSight to focus on what the market really needs rather than generating finer and finer categories to distinguish different types of content. The core element of a database filter is the number of URLs in its database, and therefore, vendors like to draw attention to the size of the database and its multitude of categories as a measure of its worth. For example, some databases make distinctions between "nudity" and "adult material" in order to "more accurately" categorize content according to differing opinions.

PureSight has found that most users and organizations do not want to make such fine distinctions. In fact, administering so many categories is more of a management challenge than a benefit. Therefore, PureSight maintains the basic categories under which any site can be accurately classified and filtered. Distinctions and decisions to block or allow access are made by how closely the content fits the basic category, and where its mathematical coordinate falls.

Bandwidth issues

One of the most important benefits of deploying a filter in an organization's network should be the reduction on bandwidth consumption. Efficient use and management of

bandwidth can reduce company expenses considerably. PureSight is able to filter content based purely on file type. So, for example, if an organization finds that its bandwidth is over-subscribed by employees downloading music files, all popular audio extensions could be filtered.

6.4. ACR Scalability

PureSight product implementation allows maximum utilization of the hardware. The product will scale in parallel with the hardware. Adding a second CPU will increase performance by almost 100% and the faster the Internet connection, the faster the response time. Furthermore, PureSight product infrastructure allows deployment of an array of PureSight servers to scale the solution to an unlimited amount of users.

6.5. Cache categorization

The ACR engine utilizes a one-pass inspection technology that analyzes packets upon arrival rather than waiting for the entire page to be downloaded, so that processing is fast and with no noticeable latency to the end user.
In addition, the PureSight solution incorporates a URL cache to store classifications of previously requested sites that were analyzed by the ACR engine. The URL cache increases performance and reduces latency by actually building a dynamic database that accurately reflects the surfing habits of the organization. The cache mechanism allows PureSight to deliver the efficiency of a database solution, yet at the same time, further reduces the minimal latency by excluding irrelevant sites from the cache.

6.6. Database Categorization

Combining the best of two worlds with a hybrid approach, PureSight includes a Static URL database. In certain cases, the ACR engine may fail to accurately categorize a page. This could happen because there is not enough information available for analysis, or because the category in question does not easily lend itself to automatic analysis. In order to effectively handle these instances, a layer of manually categorized sites is included in the PureSight products. The database is automatically updated as needed (or as requested).

6.7. Customized Site Lists

PureSight provides an interface for readily handling network specific lists of sites that can be allowed (white lists) or disallowed (black lists). These lists can be implemented in the solution as general lists or per specific user profile.

The customized lists are atop all layers and can be set to override the static URL database as well as the ACR, dynamic filtering layer. There are two methods for entering a new entry – per specific URL / page, or per a domain / site that covers all pages included within the domain.
The network administrator can also fine-tune and override the ACR. URLs and domains can be added to the different categories, or it can instruct the ACR to ignore URLs and domains, in the different categories.

6.8 Tagging URL – Rating Systems

PureSight has the ability to analyze both the request for the html page as well as the actual result of the request. Most available solutions implement the content analysis only on the request – by analyzing the URL requested against a static URL database. However, in order to support the PICS rating system that is currently supported by some web sites, a filtering solution must be able to analyze the requested page to determine the rating.

PureSight includes a PICS rating system parser. The parser analyzes a URL's meta-tags and determines the classification of the request.

PureSight supports the iCRA (formerly RSACi) and SafeSurf rating systems. These two rating systems provide a classification method based on categories such as – Adult, Gambling, Chat, Drugs etc along with Language (Explicit sexual, Crude/profanity, Mild expletives) and Age (All Ages, Older Children, Teens, Adult Supervision Recommended, Explicitly for Adults etc).

7. Summary

Internet content filtering has become an essential component in any organization's Internet security infrastructure. Businesses, schools and residential Internet users will increasingly require filtering solutions that can deliver the necessary results:

• Accurate filtering of inappropriate content while permitting all other content deemed appropriate in the organization's Internet usage policy.

• Measurable reduction of bandwidth consumption applications that are not mission-critical.

• Reduction of legal liability that could result from the existence of illegal content on the organization's network.

URL database methodology was a suitable solution when the Internet was smaller and less chaotic. Today, we see that manually mining and classifying Internet content is an impossible task. The effort to keep up requires tremendous resources but is doomed to fall far short of its goal. The numbers tell the story of what is happening. And the Internet only gets more chaotic as it grows.

The PureSight filtering layers constitute a unique multi-layer approach to Internet content filtering. It maximizes the solution's capability to accurately filter virtually all inappropriate information.

As we have explained in this paper, a multilayer dynamic filter ensures full coverage of all Web content, whether it is the most well know pornography site or some homemade pedophilia site that no search engine will ever index. A good understanding of the state of the Internet today must lead one to conclude that a truly effective filtering solution must be able to analyze content and make accurate decisions on the fly.

Through the development of the ACR engine, PureSight's engineers have succeeded in addressing the most problematic issues facing the Internet filtering solution market today by providing proven accuracy, immunity to workarounds, on the fly identification, multilingual capability, always up to date coverage, scalable capacity, and high performance.

Sources

Bergman, Michael K. "The Deep Web: Surfacing Hidden Value." BrightPlanet and The Journal of Electronic Publishing, University of Michigan Press, August 2001: Volume 7, Issue 1. (http://brightplanet.com/deepcontent/tutorials/DeepWeb/index.asp)

Brin, Sergey and Page, Lawrence. "The Anatomy of a Large-Scale Hypertextual Web Search Engine." Computer Science Department, Stanford University, Stanford CA.

Brunton, Michael. "Illuminating the Web." Time Europe, July 9, 2001, Vol. 158, no.2, (http://www.time.com/time/europe/biz/magazine/0,9868,166169,00.html)

Elkin, Toby. "At-Work Internet Users Biggest Online Spenders." AdAge.com Interactive News, September 24, 2002. (http://www.adage.com/news.cms?newsId=36121)

Frost & Sullivan. *Content Filtering Markets*, Report, 2001, pg. 11.

"How Much Information." University of California at Berkely, 2000. (http://www.sims.berkeley.edu/research/projects/how-much-info/internet.html)

Internet Software Consortium, *Internet Domain Survey*, July 2001, July 2002 and July 2004. (http://www.isc.org).

Kanaley, Reid. *Knight Ridder News Service*, "Domain Claim: Entrepreneurs are Gobbling up expired Web addresses." APLUS.NET, September 24, 2001, (http://www.aplus.lycos.com/services/domain-news-9-24-01.html)

Murray, Brian H. and Moore, Alvin. *Sizing the Internet*, *a White Paper*, Cyveillance, July 10, 2000. (http://www.cyveillance.com/web/downloads/Sizing_the_Internet.pdf)

Penn, Jonathan. "Evaluation Criteria for Web Content Security Solutions." IdeaByte, Giga Information Group, July 10, 2001.

Penn, Jonathan. *Web Content Security Needs to Evolve*, Giga Information Group, May 9, 2001.

"Staff misuse web despite policies." Personnel Today, July 9, 2002. (http://www.personneltoday.com/pt_news/news_feat_det.asp?liArticleID=13716)

Ward, Mark. "Net Filters Fail the Children." BBC News Online, March 26, 2002.

"Web abuse main reason for disciplinary action." Reuters, September 2, 2002. IT

"Firms tackle web abuse." IT Week, November 2003. (http://www.itweek.co.uk/news/1148711)

(http://www.reuters.com/news_article.jhtml?type=internetnews&StoryID=1400370)

Google Blog, November 10, 2004. (http://www.google.com/googleblog/2004/11/googles-index-nearly-doubles.html)

Top Ten Reviews, November 2004. (http://www.internetfilterreview.com/internet-pornography-statistics.html)

