Whitepaper: Understanding Web Filtering Technologies

ABSTRACT

The Internet is now a huge resource of information and plays an increasingly important role in business and education. However, without adequate controls in place, organisations are likely to be faced with a broad range of issues. These range from excessive personal use of the Internet during business impacting staff productivity to legal risks if users access inappropriate content.

This has led to the emergence of web filtering products, which organisations can deploy to enable proactive management of Internet access for users.

This whitepaper discusses some of the technology options available for web filtering and reviews the benefits and limitations of each.

WHY IS WEB FILTERING REQUIRED?

The term web filtering is broadly used to describe the process and tools that companies can use to restrict or monitor their users' Internet use.

The Internet provides an extremely effective way for organisations to increase productivity, lower costs and increase sales. It provides a quick and efficient way to undertake research, improve learning, interact with customers and transact business.

However, it is also a source of content that is very likely to be inappropriate for users to access, such as pornography, violence and racism to name just a few. In addition to inappropriate content, there are a range of other types of content, such as shopping, social networking and news that can have a huge productivity and consequently a financial impact on a company.

In addition, there a number of legal risks such as Employment and Sex Discrimination legislation, which companies may encounter if they do not proactively manage Internet access. In some circumstances this can lead to financial and reputation losses[1] .

THE EMERGENCE OF WEB FILTERING

When the Internet consisted of a few hundred thousand web sites, then web filtering was relatively simple. IT management simply had to maintain a list of "good" and "bad" sites that users were allowed or not allowed to visit. This was usually done within a network element, i.e. a Firewall.

However, the Internet quickly became so large that this approach could no longer be sustained. It's worth noting that by the end of 2007, the Internet grew by 48%[2] and the major search engines indexed 46 billion web pages[3] .

It also highlighted another issue of how to manage access to sites that may be good or bad. For example a travel site could be "good" in a business context for some users but "bad" for others.

Web filtering quickly became established due to these two key drivers.

The remainder of this white paper discusses some of the alternative technologies used in web filtering.

BASIC WEB FILTERING - FIREWALLS

The most basic level of web filtering can be performed by a network Firewall. Whilst this allows a degree of filtering, there are significant performance implications as the Firewall needs to inspect the traffic to identify the requested site to make the decision on allowing or blocking the site.

In addition, the reporting capability of a Firewall is very basic. It allows limited management review of web access, albeit after access to the website has occurred. However, this is typically labour intensive and inadequate in providing substantive proof that a particular user was responsible for any inappropriate access.

Black lists can be employed to list undesirable web addresses and prevent access to those sites. White lists can be used to list acceptable web addresses and are often used to restrict access to only those sites that are contained on the white list. The scale of the Internet is now such that maintenance of lists is a formidable challenge; for users it is very frustrating if they have a genuine reason or need to access a site but first must seek approval and have it included in allowed URLs.

In more sophisticated solutions, black and white lists are used to list exceptions to the rule for users. For example, a user may not be allowed access to travel sites but is provided access to low cost airline websites to book flights.

[1] If you would like more information about why Web Filtering is required and a more detailed discussion around the legal issues, please refer to the document "Guide to Web Access Management", available at www.bloxx.com.

[2] Netcraft

[3] www.worldwidewebsize.com

URL DATABASE WEB FILTERING

One of the most prevalent methods of web filtering is to create a database of many millions of web addresses categorised according to their content, e.g. shopping, gambling, violence, etc. Typically known as a URL database, these have increased in size over time to try to match the growth of the Internet.

The URL database can be used to effectively create a range of user Internet access profiles so that different groups of users get controlled access to the Internet.

In practise, this means that one user may get no access to the Internet, another user may only have access to a single web site, a secretary may have access to travel sites but not shopping and a manager may have wider Internet access with the exception of inappropriate sites such as violence, racism and pornography.

The demand for this type of approach has created a growing number of URL database builders who typically invest heavily to maintain and grow their URL lists.

CREATING AND MAINTAINING URL DATABASES

Many thousands of URLs are initially "harvested" each day using a variety of methods, but manual categorisation using human URL reviewers is still predominantly used to ensure accuracy of categorisation.

Manual categorisation requires each URL reviewer to read the content and look at images on the website, decide the kind of web site it is and categorise it in the database accordingly. The accuracy of this approach is variable - it is relatively easy to spot a pornography web site for example, but less easy to identify anonymous proxy sites.

CHALLENGES WITH URL DATABASES

URL databases present a number of challenges.

Misclassification With limited time available to categorise each site, any web site that deliberately seeks to mislead a reviewer (e.g. a cookery site which when examined in more depth turns out to be pornography) can easily be successful in having the inappropriate URL categorised as legitimate.

Misclassifications of web sites are extremely frustrating for users and are often a source of conflict between suppliers and customers.

Keeping pace with the growth and dynamic nature of the Internet The Internet is said to be currently growing by approximately 7.5 million new or renamed web addresses each day. However, a URL classifier will typically review and classify only around 500 web addresses daily. To keep pace with this growth and to keep a URL databases current and relevant would require around 15,000 classifiers. From a cost perspective this is clearly unrealistic.

In addition, there is a lack of protection against zero-day threats from newly created web sites as it may take several weeks or months before the site is harvested, reviewed and added to the database.

In addition web filtering suppliers are not particularly motivated to re-check URLs previously categorised. A significant volume of sites are either re-named or cease to be operational, and to delete these from URL lists risks stopping them from promoting an ever growing database. Therefore the claims of web filtering suppliers about database sizes require further inspection to reach any conclusion about their capability.

Scale Companies who provide web filtering usually try to differentiate on the basis of the size of their database of categorised sites - it is usual for these databases to contain 15 to 35 million categorised sites. The principle they conform to that the larger the URL database then the better the coverage and protection offered

However, in the context of the overall size of the Internet, the URL database, no matter its size is inconsequential and means that IT managers need to decide if they will allow or deny access to requested URLs not listed in the database.

Allowing access to any requested URLs not listed in the URL database means that users have open and uncontrolled access to a vast part of the Internet which will only exacerbate the risks and dangers previously noted. This underblocking is embarrassing for suppliers, but also, in certain markets, i.e. education, is completely unacceptable.

Taking the opposite approach and blocking all requested URLs not listed in the database results in overblocking, and usually means an increase in user frustration and an increase in IT management time spent adding sites to a white list.

IMAGE SCANNING

Image scanning is a useful but not a foolproof way of blocking pornography from a network. It tends to be expensive and processing intensive and is not a realistic solution for managing Internet access.

URL KEYWORD SCANNING AND SCORING

URL Keyword Scanning and Scoring is typically used to complement URL Database web filtering to provide an additional layer of protection. Keyword Scanning examines the keywords requested by the user either within the web address requested. These vary in sophistication where used.

IN-LINE PAGE KEYWORD SCANNING

Similar to URL keyword scanning and scoring, In-Line Page Keyword Scanning operates by scanning the text on a requested web page before it is delivered to a user. Lists of keywords and offensive words are created and given positive or negative scores. When users request a page, the content is analysed for occurrences of these keywords and if it exceeds the scoring threshold set for the user, then the page is blocked.

Some suppliers allow IT management to tailor keyword requests to ensure vertical market needs are addressed. An example of this would be a building college where a basic system stops students searching for wire strippers, road hardcore etc. These systems can be configured to allow these specific phrases but prevent the searching of the word “strippers” alone.

This type of filtering has its uses for offensive content plus it does not need to have previously identified the web site and categorised it to block it. It also solves some of the issues concerning false positives (e.g. place names).

However, also In-Line Page Scanning is only really of benefit when blocking some offensive sites, as these have a distinct vocabulary. It will not, for example, be able to differentiate whether a sports story is on a sports site or a news site. Even then, if offensive words from different languages are not listed, sites containing this content will be allowed.

This method does not take context into consideration and often blocks sites that users should be allowed access to. For example, Keyword Scanning tends to only be able to differentiate a gambling web site from a “help from gambling” web site if “good” keyword combinations are listed and negate the score gained by the offensive word. If “good” keywords are not listed or used sufficiently often, perfectly reasonable sites (e.g. Sex Education) can be wrongly blocked.

REAL-TIME CONTEXTUAL ANALYSIS AND CATEGORISATION

Currently, most web filtering suppliers use variations and combinations of the techniques previously described. However, web filtering suppliers are now starting to realise that these filtering techniques alone cannot cope with the growth and dynamic nature of the web. Millions of web pages are created every day and are simply missed by web filters because they haven't yet been added to its URL database or are not picked up by the URL/page scanner.

New technology is now emerging using advanced software techniques to analyse the patterns and context of the text on a page at the point the page is requested and before it is presented to the user.

Bloxx Tru-View Technology is an example of this type of approach to web filtering.

The software operates by using language patterns and contextual information previously gathered from a number of web pages across a number of specified categories, and then analysing the requested page to automatically categorise the page based on matching signatures.

This in-line, real-time method of web filtering allows web pages that have never been discovered before and consequently are not yet listed in the URL database, to be identified and categorised correctly with an extremely high level of categorisation accuracy.

The method is extremely effective at categorising web pages across a wide range of different categories, not just inappropriate content such as pornography. For example, the software is excellent at categorising content such as shopping and social networking that may not be inappropriate but could have a dramatic impact on staff productivity.

When used in conjunction with some of the existing web filtering approaches such as a URL database and Keyword Scanning, this multi-layered approach to web filtering provides an extremely effective method that copes well with the demands of today's Internet.

In a recent Technology Audit from industry analyst Butler[4] , the firm talked about this being the third generation of web filtering, and confirmed that existing web filtering technologies were no longer able to cope with the growth of the Internet and offer limited protection to organisations and their users.

Summarising the Bloxx approach, Butler Analysts Andy Kellett commented:

"In Butler Groups opinion, the multi-layered protection and contextual analysis approach to Web filtering used by Bloxx in its TVT solution provides a highly-competitive, leading-edge solution in a marketplace that is often driven by vendors whose web filtering solutions still rely more heavily on the frantic need to keep URL databases up-to-date."

SUMMARY

URL Database web filters are falling further behind as a result of rate of growth in the Internet every day. No supplier can invest in enough resources to match Internet growth and amendment rates using this method alone.

Black and White lists, URL lists, Keyword Scanning all depend upon previously scoped defence mechanisms - where a web site does not fit in with previously scoped words, phrases, exceptions etc, the web filtering fails to be effective.

Keyword page scanning is most effective in web sites with an emphasis on unique language for a particular type of site - e.g. slang words on porn sites. The ability to identify mainstream web sites by keyword scanning is very much more limited and in some cases, completely ineffective.

Some web filtering solutions are not capable of any form of page content analysis due to the implementation methods used. These solutions are forced to concentrate only on large URL databases to provide the filtering.

Real time contextual analysis and categorisation can be used in conjunction with other web filtering techniques to provide a web filter that copes most effectively and efficiently with the demands of today's Internet.

[4] A copy of the Butler Group Technology Audit on Bloxx Tru-View Technology can be found at www.bloxx.com.

ABOUT BLOXX

Headquartered in the UK with sales offices in Australia, Netherlands and the USA, Bloxx offers web filtering appliance-based solutions for medium and large organisations in both the business and public sectors. Leading UK investment groups such as Braveheart Investment Group Plc and Archangel Investments Ltd. have invested in Bloxx. For more information please visit: **www.bloxx.com.**

ABOUT BLOXX TRU-VIEW TECHNOLOGY

Bloxx Tru-View Technology uses internationally patent pending technology to analyse and block web sites quicker and more accurately than other web filters, which use manual classification and keyword scoring. Tru-View Technology uses intelligent identification and analysis providing instant classification of web content as soon as it is accessed even if the content has not been seen by anyone before.

Bloxx Tru-View Technology helps organisations proactively manage users' access to web content which might lower productivity, expose the organisation to risk and liability or pose a network security threat.

An estimated 1 million + users already benefit from enhanced security and performance with low administration and no cost per user charges. Additional protection is provided via anti-virus, anti-spyware and anti-phishing functionality, alongside onboard cache.

t. 800 876.3507 e. sales@journeyed.com w. www.journeyed.com