IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP4143

The Effectiveness Of A Web-Filter on Domain Specific Users

Geoffrey Sandy and Paul Darbyshire
School of Information Systems, Victoria University, Melbourne City MC, Australia, {Geoff.Sandy, Paul.Darbyshire}@vu.edu.au

ABSTRACT

As the amount of content on the Web grows almost exponentially, one of the new growth industries is that of filtering products. The effectiveness of web-filtering software depends on a number of factors including the architecture of the software itself, and the sophistication of the users operating within its application domain. The main use of filtering software is to "block" access to controversial content such as pornography. This paper reports an investigation of the effectiveness of a filter called squidGuard in the real-world environment of an Australian University. The product is used to "block" pornographic material. This investigation simulates two classes of web users in trying to access pornography. While squidGuard did have limited success in blocking such material from novice users, the blocking rate dropped dramatically for the more experienced users using access lists. In all cases however, access to supposedly filtered material was gained in seconds. Under such testing, the effectiveness of squidGuard as a specific-content filter for "pornographic" material can only be seen as superficial approach at best.

INTRODUCTION

Filter software is increasingly used by a wide variety of groups in society and in many societies its use is mandated by law. Filter software is used in the home and school markets. Parents and teachers use a filter to prevent children from accessing content deemed not suitable for them. Sexually explicit and violent material is of most concern to parents and teachers. The regulation of controversial material in respect to children is also an issue for organizations like libraries and universities. Recently the corporate world has embraced filter technology because of concerns expressed about Internet content.

Fundamental to the acceptance and use of a filter are two questions. First, how effective is it in blocking content that is intended to be blocked. Second, how effective is it in not blocking content that is intended not to be blocked. Vendors claim their product is highly effective. Many vendors also claim that the product is highly effective because before content is blocked it is evaluated by a person using a rating system.

This paper reports on testing the effectiveness of a filter product called squidGuard that is used in a number of Australian universities. The test is conducted in a real-world environment at one of these Universities (referred to as University X), and simulates different types of consumers of Internet pornography that may be found in this environment. University X mainly uses the filter to block what the vendors blacklist describes as pornography. The University does add it own sites to the blacklist and the product offers a number of blacklists additional to pornography.

In the following sections some background material is provided on the effectiveness of other filtering products, and a description of the squidGuard filter is provided. A classification scheme is devised which classifies users that browse the Web for Internet pornography in terms of their sophistication in browsing for specific-content. A first set of trials is described that test the effectiveness of squidGuard in blocking material intended to be blocked for two levels of users. A second set of trials is described to test the effectiveness of squidGuard in blocking material that is not intended to be blocked. The results of both sets of trials are presented, and discussed and some areas for future research are explored.

BACKGROUND

Filter software is designed to block access to controversial Internet content. Such content can be blocked in three places: at the source, that is the provider or creator of the content; in transit either at the application level or at the packet level; at the receiver. Blocking at the receiver end may be done directly by installation of a filter on the receivers PC, or indirectly at the receiver's Internet Service Provider (ISP) with a subscription to a "whitelist". The process of blocking requires the content must be rated based on some classification system. Such a system may be simple which produces a rating of allowed/ disallowed or a sophisticated one like the Platform for Internet Content Selection (PICS) that employs many categories and values. Opinion is divided over whether PICS or similar labelling systems is a solution to blockng controversial content, see for instance (Kohntopp and Kohntopp 1999; Chen et al 1999; Hochheiser 1997).

There is a growing body of research that challenges the effectiveness claims of filter software vendors. Haselton (2000a) using "zone files" from Network Solutions (which lists all .com files) obtained a list of the first 1000 active .com domains for June 14 2000. This list was tested using five popular filters to discover how many sites were blocked as "pornography", and of those sites how many were actually pornographic. The products tested were "Cyber Patrol" (Haselton 2000b), "SurfWatch" (Haselton 2000c), "Bess" (Haselton 2000d), "AOL Parental Controls" (Haselton 2000d) and "SafeServer" (Haselton 2000e). The error rate found for each product was computed as the number of non-pornographic sites blocked divided by the total number of sites blocked. The average error rates for each filter product were "Cyber Patrol" 81%, "SurfWatch" 82%, "Bess" 27%, "AOL" 20% and "SafeServer" 34%. The researcher believes that the claim of human evaluation of material is false and the error rate would be higher if .org sites were tested.

Haselton (2000f) tested the "BAIR" filter with a sample of 50 randomly selected pornographic images and 50 randomly selected non-pornographic images. Haselton (2000g) retested the filter on July 18 2000 and found that 34 out of the 50 pornographic images of the first experiment blocked. Of the non-pornographic images only 8 were blocked. A random selection of 50 images of peoples faces were selected to test if they were blocked. Out of 50 face images 34 were blocked. It was concluded that "BAIR" has only negible ability to distinguish between pornographic images and pictures of peoples faces.

Finklestein (2000) investigated "SmartFilter". The study provides empirical eveidence to confirm the mathematical impossibility of a human evaluation of blacklisted content. It lists many sites blocked by the product that was not intended to be blocked. Possible programming-related reasons are put forth as to why these sites are blacklisted.

The Censorware Project (2000) tested the effectiveness of "Bess" . Thousands of URLs were tested against "Bess" proxies in real-world use from 23 July to 26 July 2000. The ten proxies were configured similarly to each other and to the setup that "Bess" recommends for schools. The major finding is the "Bess" is ineffective in blocking many porn sites, and mistakenly blocks a great deal of useful non-pornographic material suitable for school children. A test was also made by The Censorware Project (1999) of "SmartFilter" used by the Utah Education Network that resulted in similar findings.

This paper appears in *Issues and Trends of Information Technology Management in Contemporary Organizations,* the proceedings of the Information Resources Management Association International Conference.

An independent review was conducted of the "Clairview Internet Sheriff" in May 1999 (Electronic Frontiers Australia 1999). Internet access accounts were purchased with Clairview's Brisbane ISP, Cvue Internet. Cvue start-up packages include 20 hours access and customers using this service are anonymous to Clairview. The customers (three reviewers) simulated the use of the internet as a cautious and reasonably internet familiar parent would to check the effectiveness of the filter for themselves and children. Approximately 20 hours of testing was conducted by three people for a week. Access was sought to sites unambiguously pornographic. Many pages were accessible. On returning to these pages later some were now blocked. It was found the blocking mechanism was able to be by-passed using free anonymiser-type services available on the World Wide Web, that is, blocked sites could be accessed while using the Cvue service. It was also found to block vast numbers of non-pornographic pages.

Peacefire, a Youth Alliance Against Internet Censorship over a number of years have published reports on sites and newsgroups that could not be described as pornographic but have been blocked by filters. Online. These include "Cyber Patrol", "Net Nanny", "X-Stop" and "CYBERsitter".

The Electronic Privacy Information Centre (EPIC 1997) conducted 100 searches using a traditional search engine and 100 searches using a search engine described as the "world's first family Internet search site". An attempt was to locate material that might be useful to children. This included schools and charitable, educational, artistic and cultural institutions. Search terms included "Smithsonian Institute", "American Red Cross" and San Diego Zoo" It was found that the family-friendly search engine prevented access to 90% of materials available on the Net using the relevant search terms. It was also found that the family-friendly service denied access to 95%-99% of material otherwise available without filters. The study concluded that the filtering mechanism prevented children from accessing appropriate material likely to be useful to them.

FILTER DESCRIPTION

Information Technology management of University X state that squidGuard was introduced for two main reasons. First, to minimise risks of litigation and the possible infringement of sexual harassment legislation. Second, to contain Internet costs. Notwithstanding this, squidGuard is universally applied across University X and applies to both staff and students.

squidGuard blocks sites at the application level by filtering material from the user based on the destination of the URL requests. Sites can be filtered based on the following:

- User ID
- IP Addresses
- Entire domains, including sub-domains (my.domain.com)
- Entire hosts (Host.my.domain.com)
- Directory URL's (my.domain.com/directory/one)
- Specific files (my.domain.com/directory/one/file1.html)

The IP addresses, domain URL's, hosts and files to be filtered are compiled into a database and searched every time a URL request is made through the filtering server. If a match is found then the user is directed to a URL specified through the squidGuard configuration. The databases are compiled by periodically initiating a "dumb" Web robot[1] to scan Internet URL's based on keywords and expression-lists to find Web pages to block. The robot consumes large amounts of computer resources so it is recommended that it be used sparingly, or that users of squidGuard contribute to common blacklists that can be downloaded from the Internet.

Web pages are not checked for content, and in fact squidGuard makes the following disclaimer:

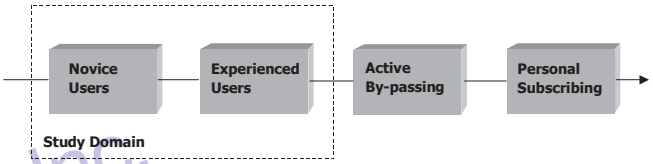
> "*The blacklists are entierly products of a dumb robot. We strongly recommend that you review the lists before using them. Don't blame us if there are mistakes, but please report errors with the online tool*".

The size and structure of the blacklist however, particularly the one concerned with pornography is such that a manual check, while not impossible, is certainly not feasible.

USER CHARACTERISTICS

In any organisation containing computer-based technology, there will be varying levels of user sophistication. When filtering software is in place whose purpose is to block entry of users to specific sites, or a class of sites, it will have varying degrees of success depending on the sophistication of the technology users. In this study we identify four levels of users with an increasing level of sophistication in browsing for specific content on the Web. These user-levels are indicated in Figure 1, and represent an increasing sophistication from left to right in the diagram.

Figure 1: ***User sophistication and domain of study***



There are four classes identified in Figure 1.

Novice User- someone who wishes to access filtered material on the Web but has no great experience in doing so. Such a user in all likelihood would begin with a standard search engine using obvious keywords from their learned experience.

Experienced User - users who have been successful in previous browsing, and have gained some experience in searching for the specific content. Such a user would be familiar with less obvious keywords to use in standard search engines, or perhaps more likely, follow links on daily updated specific content lists.

Active By-passers - some users, aware of the filtering technology and techniques they employ, attempt to take active steps to bypass the filtering technology. There are many Web sites devoted to instruction on techniques to use.

Personal Subscribers - this class of user is represented by people who personally subscribe to filtered content by filling out Web forms, and subscription lists. The filtered material can then be delivered to their personal mailboxes, sometimes in innocuous format.

The last two classes of user are difficult to dissuade, and quite often the only way to prevent their access is to remove them from the Web altogether. The first two classes of users represent our domain of study. We believe that users within this domain represent people who may be willing to browse the Web for specific content, but may be discouraged depending on the success of the filtering software

DATA COLLECTION

The Data collected during this research was designed around the simulation of the two classes of user in our study domain trying to gain access to filtered pornographic material.

In simulating the *Novice User*, two Web search engines were chosen, Google,[2] and Alta-Vista.[3] For both of these search engines, 10 initial trials were conducted. A trial was conducted for each of 10 keywords that is 10 Web searches of filtered material. In each of the 10 searches, the first 20 URL's were followed to try and gain access to the filtered material. This represents 200 URL's per search engine. For each search, a tally was kept as to the number of successful accesses (accesses to filtered material) and the number of sites filtered (material was successfully filtered). In some cases, errors may have occurred due to problems with the URL's, and these were also noted. The percentages of the *access rate* and *filtered rate* were than calculated and tallied.

The 10 keywords used for the trials were 10 obvious *porn-keywords*, chosen from an industry compiled list of unambiguous pornographic terms. The term "*obvious*" is subjective, but the keywords were chosen from the list for two reasons. The keywords were deemed to be socially familiar (if not acceptable) to the wider community, and Web sites indexed using these keywords would contain unambiguous pornography. Unambiguous pornography is material that would be designated as pornographic in any broad-based definition.

A second trial was then conducted, similar to the one just described, except a different set of 10 keywords were selected. These keywords were selected as they were *less-obvious* terms to use when searching. Again, the term "*less-obvious*" is subjective, but they were chosen from the same industry compiled list of unambiguous terms. Such terms represent search criteria used by more experienced users of pornography.

Finally, a trial was conducted to simulate a more *Experienced User* by using comprehensive, daily-updated, specific content lists. These lists are readily accessible via the Internet, and most users browsing for pornographic content will come across them, as the researchers did during the first trials discussed above. Two common lists were chosen and for each list the first 50 URLs were checked for accessibility.

A trial was also undertaken to search the Web using keywords such as "*sexuality*", "*gay*", "*lesbian*" and "*sexual health*". In each case, the first few links were attempted. If these were blocked, then their content was checked against an un-filtered Internet connection through an external ISP. This test was performed to gauge the extent to which the filter inappropriately blocked access to sites where the content was unambiguously, not pornographic in nature. The keywords and lists used are not presented in this paper as they may give offence to some readers. Details are available on request from either of the authors.

RESULTS

In the first trial of checking two common search engines for 10 "obvious" keywords, a table is presented that represent the results from both search engines. Table 1 represents the summarized results of the trial using the Google and AltaVista search engines. In the case of an error (page inaccessible), the page was not counted in the resultant statistics.

The summarized results for the two corresponding trials of the 10 "*less-obvious*" keywords for the Google and Alta-Vista search engines can be seen in Table 2.

The results shown in Table 3, represent the simulation of an experienced user using content specific lists which are updated on a daily basis.

In the trials for testing the filter inappropriately blocking non-pornographic material, 26 sites were identified as containing non-pornographic material and were blocked by squidGuard. At this point the researchers stopped, as it seemed clear that many more could be identified if the trial persisted. Space does not permit detailing the nature of all the sites here but many of these sites included material on:

- Safe sex practices
- Planned parenthood
- Alternative lifestyles
- Sexually transmitted diseases including HIV

One site blocked was also a Professional Journal, "Journal of Sexuality and Culture".

Table 1: Trial 1, 10 obvious keyboards with Google and AltaVista search engines

		Google		AltaVista	
Keyword		**Access Rate%**	**Filter Rate%**	**Access Rate%**	**Filter Rate%**
1		58	40	16	84
2		20	80	25	75
3		30	70	35	65
4		40	60	15	85
5		47	53	20	80
6		42	57	10	90
7		32	68	35	65
8		42	58	20	80
9		35	65	5	95
10		55	45	25	75
	Average%	**40%**	**60%**	**21%**	**79%**

Table 2: Trial 2, 10 less-obvious keywords with Google and AltaVista search engines

		Google		AltaVista	
Keyword		**Access Rate%**	**Filter Rate%**	**Access Rate%**	**Filter Rate%**
1		15	85	35	65
2		35	65	35	65
3		70	30	40	60
4		20	80	10	90
5		30	70	25	75
6		35	65	25	75
7		25	75	20	80
8		75	25	5	95
9		70	30	20	80
10		45	55	25	75
	Average%	**42%**	**58%**	**24%**	**76%**

Table 3: Trial 3, using pornographic lists

List	**Successfully Accessed**	**Filtered**	**Errors**	**Access Rate%**	**Filter Rate%**
List 1	37	13		74	26
List 2	38	12		76	24
		Average %		**75%**	**25%**

DISCUSSION

The data presented in the Results section indicates a varying degree of success of the filtering mechanism in blocking the intended specific content. The filter rate varied quite widely depending on the keyword used to search, even within the same search engine. Generally speaking, the filtering software was more successful at blocking specific content when the more obvious keywords were used, as opposed to the less obvious ones. The differences between the *obvious* and *less-obvious* keyword results were not significant.

There was a significant difference however between the filter rates of the specific content lists and the standard search engine results. The highest average filter rate achieved using the standard search engines was with Alta Vista at 79% using the *obvious* keywords. But the lowest filter rate achieved by using the two lists was 58%. The use of daily updated specific-content lists represents a more experienced user according to our hierarchy of user sophistication in Figure 1. Thus the more sophisticated users are able to gain access to a greater number of specific content sites. As the lists are updated on a daily basis, it is more probable that many of the new URL entries are actually new to the Web, and hence not already filtered by such products as squidGuard. Thus a lower filter rate could have been anticipated as proved to be the case.

There are a number of important issues that relate to filter products like squidGuard. First, is a recognition that they are highly inaccurate in that they fail to block targeted material, and material is blocked that should not have been. Users are unable to verify the inaccuracy as access is denied to the material. The reality concerning effectiveness of the product is very different from the claims made by vendors.

The vast size of available material on the Web, it's growth rate (in Volume), and the frequent changes made to existing sites makes evaluation very difficult. Although vendors claim that sites are evaluated by humans before being blocked, the reality is that vendors largely

use computers for evaluation. The vendors of squidGuard honestly inform users of the product that its blacklists are entirely the results of a dumb robot. This is inappropriate as anybody not checking the lists manually could be filtering out invaluable material. Much of the URL's provided in black-lists are the results of personal judgements about the information requirements of Internet users made by anonymous third parties. There is a risk that these third parties adopt a very conservative attitude to material, or worse incorporate their prejudices and biases.

CONCLUSION

In this paper we have investigated the success of filtering software, in particular squidGuard, in filtering specific-content from Web browsers. Four levels of Internet users have been identified with increasing sophistication in searching for undesired content. As a users level of sophistication increases, their chances of gaining access to this material will increase dramatically.

Three trials were conducted to test the effectiveness of squidGuard on two levels of user sophistication. This was done by simulating possible access methods by these users. In the first trial, two standard search engines were used to test access success when users searched using obvious keywords. The second trial was similar except less obvious keywords were used with the search engines. In the third trial, lists containing URL links to the specific-content were used. These lists are readily available on the Web, but it takes some time to find them. Thus, only more sophisticated browsers would use such lists rather than a standard search engine.

While the squidGuard did have limited success in blocking material from a *Novice User*, the blocking rate dropped dramatically for the more *Experienced User* who used the access lists. In all cases however, access to supposedly filtered material was gained in seconds. Under such testing, the effectiveness of squidGuard as a specific-content filter for "pornographic" material can only be seen as superficial approach at best.

ENDNOTES

1 http://ftp.ost.eltele.no/pub/www/proxy/squidGuard/contrib/squidGuardRobot/

2 http://www.google.com/

3 http://www.av.com/

REFERENCES

Chen D et al (1999) *Centralized Content-Based Web Filtering and Blocking: How Far Can it Go?* IEEE, pp115-119.

Electronic Frontiers Australia Inc. (1999) *Report: Clairview Internet Sheriff An Independent Review*, http://www.efa.org.au/Publish/report_isheriff.html, 24/10/2000.

Finklestein S (2000) *SmartFilter – I've Got A Little List: An anticensorware investigation*, http://sethf.com/anticensorware/smartfilter/gotalist.php, 2/3/2001.

Haselton B (2000a) *Study of Average Error Rates for Censorware Programs*, http://www.peacefire.org/error-rates/, 25/10/2000.

Haselton B (2000b) *Cyber Patrol error rate for 1,000 .com domains*, http://www.peacefire.org/censorwa...atrol/first-1000-com-domains.html, 25/10/2000.

Haselton B (2000c) *SurfWatch error rate for first 1,000 .com domains*, http://www.peacefire.org/censorwa…Watch/first-1000-com-domains/html, 25/10/2000.

Haselton B (2000d) *Bess error rate for 1,000 .com domains*, http://www.peacefire.org/censorware/BESS/second-1000-com-domains.html, 25/10/2000.

Haselton B (2000e) *SafeServer error rate for first 1,000 .com domains*, http://www.peacefire.org/censorwa…Proof/first –1000-com-domains.html, 25/10/2000.

Haselton B (2000f) *BAIR "image filtering" has 0% accuracy rate*, http://peacefire.org/censorware/BAIR/first-report.6-6-2000.html, 5/5/2001.

Haselton B (2000g) *BAIR cannot distinguisk between pictures of faces and pornographic images*, http://peacefire.org/censorware/BAIR/second-report.7-19-2000.html, 2/3/2001.

Hochheiser H (1997) *Filtering FAQ*, http://www.eff.org/pub/Censorship/Rati…ters_labelling/HTML/filtering_faq.html date accessed 10 December 2000.

Kohntopp K and Kohntopp M (1999) *Content Rating and Selection does not work*, http://www.Koehntopp.de/kris/artikel/rating_does_not_work.html date accessed 20 June 2000.

The Censorware Project (2000) *Passing Porn, Banning the Bible*, http://www.censorware.org/reports/bess/, 10/1/2000.

Electronic Privacy Information Center (2000) *Mandated Mediocrity: Blocking Software Gets a Failing Grade*, http://www.peacefire.org/censorware/BESS/MM/, 25/10/2000.

Sandy G (2000) "Censorship of the Internet: Theories and Evidence of Pornographic Harm", in *Second Australian Institute of Computer Ethics Conference*, (eds.) J Barlow and M Warren, Imation (CD Rom).