To Appear in Communications of the CACM, September 2004

Calculating Error Rates for Filtering Software*

Paul J. Resnick, PhD
School of Information, University of Michigan

Derek L. Hansen
School of Information, University of Michigan

Caroline R. Richardson, MD
Department of Family Medicine, University of Michigan Medical School
VA Health Services Research and Development Service

Corresponding Author:
Paul J. Resnick
School of Information
The University of Michigan
314 West Hall
Ann Arbor, MI 48109-1092

Email: presnick@umich.edu
Phone: 734-647-9458
Fax: 734-764-2475

*This research was supported under a contract with the Kaiser Family Foundation.

Surveys in the U.S. have found that 95% of schools [4], 43% of public libraries [5], and 33% of teenagers' parents [8] employ filtering software to block access to pornography and other "inappropriate" content. Many products are also now available to filter out spam email.

Filtering software, however, cannot perfectly discriminate between allowed and forbidden content, resulting in two types of errors. First, under-blocking occurs when content is not blocked that should be restricted. Second, over-blocking occurs when content is blocked that should not have been restricted. Steps can be taken to reduce the frequency of errors, and to reduce their costs (e.g., providing easy appeals processes and quick overrides and corrections) but some errors are inevitable.

The frequency of errors is an empirical question of great importance. For, example, in 2000, the U.S. Congress passed the Child Internet Protection Act (CIPA) mandating that schools and libraries install content filtering software in order to be eligible for some forms of federal funding. A district court struck down the requirement for libraries on the grounds that it violates the First Amendment. Much of that court's finding of facts was devoted to analyses of error rates [1] and some of the arguments made on appeal to the U.S. Supreme Court also hinged on analyses of error rates.

Most empirical studies of error rates have suffered from methodological flaws in sample selection, classification procedures, or implementation of blocking tests. Results have also been interpreted inappropriately, in part because there are two independent measures of over-blocking that are sometimes confused, and likewise for under-blocking. This paper presents a framework to guide the design and interpretation of evaluation studies. While the framework applies with only minor modifications to evaluation of spam filters, the examples and discussion here focus on pornography filters.

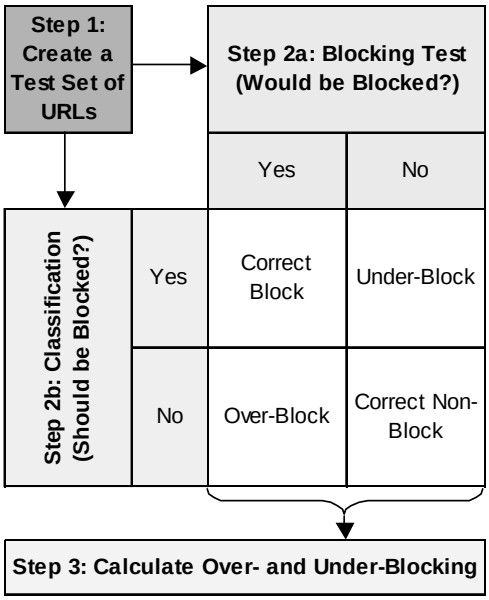

Figure 1: Summary of process for testing filter effectiveness

II. A Framework for Testing Filtering Software

The process of testing filter effectiveness is graphically outlined in Figure 1. A test set of items is generated. These items are classified to see whether they should be blocked and are tested to see whether they are actually blocked by filters. For each item, then, there are four possible outcomes: it may be correctly blocked, incorrectly blocked (which we refer to as an over-block), correctly not blocked, or incorrectly not blocked (which we refer to as an under-block). Finally, in step 3, the rates of over- and under-blocking are calculated.

Step 1: Create a Test Set

The first major step in the process is to create a test set of web sites or other Internet content on which the filters' performance will be judged. One approach is to collect a set of *accessed* items, as a way of evaluating filters' impact on users. For example, for the CIPA case, Cory Finnell selected sites from the access logs of three public libraries' proxy servers [1]. Simulations can also be conducted, to approximate what users might access. For example, for a

study of filtering error rates on health information, we entered search strings on 24 health topics into six search engines and collected the first 40 results from each search [7].

A second approach is to collect a set of *accessible* items, as a way of evaluating filters' impact on publishers. There is no good way to sample from all the available web pages (even search engines index only a fraction of pages that they encounter). Instead, some well-defined subset of Internet content must be chosen, such as the health listings from certain portal sites or all the web pages served by particular web servers.

Test sets are only representative of the larger collection from which they were drawn. For different purposes it is appropriate to estimate error rates for different subsets. For example, even within the overall domain of health sites, our study found quite different error rates from searches on the terms "condom" and "gay" than for searches on "depression" and "breast cancer" [7].

The collection process should satisfy three properties. First, it should be objective and repeatable. Many studies have relied on tester judgment to select interesting or relevant items [2, 3, 10], possibly introducing bias. Second, the collection process should be independent of the filters to be tested. The sample used by Finnell reflected patrons' access patterns when filters were installed, not what their access patterns would have been without filters. Third, large test sets should be assembled. Some studies have relied on small test sets. Others with large test sets covered so many categories of content that there was not enough statistical power to evaluate filters' effectiveness for particular categories [2, 3, 10].

Step 2a: Blocking Test

Each selected URL is tested against the various filters to see whether access to the site is blocked or not. This is best performed through automated processes that are able to quickly test a

large number of URLs against the filters. Automated tests must take into account the possibility that sites may redirect browsers, through HTTP headers, HTML, or javascript code, to other sites. A Web browser would attempt to access the original URL and then the destination URL. Thus, in an automated test, a filter should also be tested against both URLs and the site should be considered blocked if either one is blocked.

Vendors regularly update the contents of their blocking lists and rules. In order to maintain comparability between vendors, therefore, all products being compared should be updated just before the tests are run. In addition, all tests should be run simultaneously or nearly so, to allow for a fair comparison. If the test set reflects the results of simulated searches, the blocking tests should be conducted as soon as possible after the searches are run, so that the results reflect what would have been accessible to a user from the search.

Product configuration choices can have a large impact on rates of over-blocking and under-blocking. For example, nearly all products offer a variety of settings or categories that can be chosen. These categories range from pornography to gambling to hobbies and rarely match up perfectly across products, making comparisons across products difficult. Calls to 20 school systems and libraries confirmed wide variability in their configurations and that none was using a vendor's default setting [7]. Thus, tests should be run against a range of configurations.

Step 2b: Classification of Sites

Each URL in the test set is classified to determine whether it should have been blocked or not. The definition of what should be blocked will depend on the purpose of the test. For example, in order to test the over- and under-blocking of pornographic material it would be necessary to classify each site as containing or not containing pornographic material. In order to test whether filtering software implements the CIPA standard, or the legal definition of

obscenity, sites would have to be classified according to those criteria. And if the goal were simply to test whether filtering software correctly implements the vendor's advertised classification criteria, the sites would be independently classified according to those criteria.

Ideally, the classification process should satisfy three properties [6]. First, it should have face validity, meaning that there is an obvious connection to the underlying definition of what should be blocked. Second, the procedure should be reliable, meaning that the process is sufficiently documented to be repeatable and that multiple ratings of items would be in substantial agreement. Third, there should be construct and criterion validity, meaning that the classifications should be in substantial agreement with those produced by other processes that have reliability and face validity.

Because site content can change over time, sites should ideally be classified according to their state at the time that the blocking tests were run. By caching the contents of sites as of the time blocking tests are run, it is acceptable to delay the actual classification. This also allows the cache to be made public, so that others can scrutinize the classification decisions made by the raters in the study or classify the sites independently according to different criteria.

Step 3: Over- and Under-Blocking Reporting

For any product configuration and set of URLs tested, there are four results from the testing and classification, as shown in the top part of Figure 2: (a) the number of correct blocks, (b) the number of under-blocks, (c) the number of over-blocks, and (d) the number of correct non-blocks. For brevity, we will refer to sites as "bad" if they should be blocked and as "OK" if they should not be blocked according to the classification that was done: no value judgment is intended.

Figure 2: Calculating error rates

	Blocked	Unblocked
Bad	a	b
OK	c	d

Over-blocking errors

OK-sites overblock rate (1-specificity): $\frac{c}{d+c}$

Blocked-sites overblock rate (1-precision or 1- pos. pred. value): $\frac{c}{a+c}$

Under-blocking errors

Bad-sites underblock rate (1-recall or 1-sensitivity): $\frac{b}{a+b}$

Unblocked-sites underblock rate (1- neg. pred. value): $\frac{b}{d+b}$

Two-by-two outcome tables arise when evaluating all sorts of binary decisions, from radar operators detecting the presence or absence of enemies to medical diagnostic tests to information retrieval techniques that select documents from a large corpus. The most useful summaries of filtering test outcomes describe under-blocking and over-blocking error rates (i.e., percents). There are two natural ways to calculate each error rate, each providing different information. Figure 2 summarizes how to calculate the error rates and their relation to measures usually reported in information science and medical research.

Consider, first, the amount of over-blocking. One measure, which we call the *OK-sites overblock rate*, is the fraction of acceptable sites that are blocked. This measure is related to what medical researchers would call the specificity of a diagnostic test. It is useful in answering the question of how frequently a user who is trying to access OK (e.g., non-porn) sites will be

blocked. This is the number that a school or library or parent should consider when deciding whether a filter is overly broad in restricting access to information that should be available.

This error rate could also be relevant to a U.S. court performing an "intermediate scrutiny" or "reasonableness" analysis. To be reasonable, restrictions must not interfere substantially with the legitimate uses of a forum. One interpretation is that over-blocks must be few in relation to correct non-blocks of OK sites: in other words, the OK-sites overblock rate must be low.

A second measure of over-blocking, which we call the *blocked-sites overblock rate*, is the fraction of all blocked sites that are OK (e.g., not porn). This measure is related to what information scientists would call precision and medical researchers would call positive predictive value. It might be useful to a school or library or parent when deciding whether to monitor for blocks as evidence of violation of acceptable use policies. For example, if a high proportion of blocked sites are in fact OK, then the mere fact that a user tries to access a blocked site would not be a reason to suspect that user of trying to access pornography.

This error rate could also be relevant to a U.S. court performing a "strict scrutiny" analysis. To satisfy strict scrutiny, restrictions must be "narrowly tailored" to meeting a compelling government interest. One interpretation is that over-blocks must be few in relation to correct blocks of bad sites: in other words, the blocked-sites overblock rate must be low.

Note that the two measures of over-blocking are independent, as illustrated in Tables 1 and 2, which give results from hypothetical tests of two filters, on the same set of sites. In both tables, the fictitious filters have a blocked-sites overblock rate of 50%: they are equally imprecise. They differ in the OK-sites overblock rate, however. In Table 1, 99% of the OK sites are blocked but in Table 2 only 1% are blocked.

Table 1

	Blocked	Unblocked
Bad	99	1
OK	99	1

Table 2

	Blocked	Unblocked
Bad	1	99
OK	1	99

Table 3

	Blocked	Unblocked
Bad	99	1
OK	1	99

Table 4

	Blocked	Unblocked
Bad	99	1
OK	99	9801

Any estimate of the blocked-sites overblock rate is sensitive to the prevalence of OK sites in the test set. Table 4 differs from Table 3 only in having a higher concentration of OK sites. The error rates of the filter on bad and OK sites are both 1% in both tables. The blocked-site overblock rate, however, goes from 1% to 50%.

Consider, for example, Edelman's selection of 6,777 blocked sites as presented in the CIPA case [1]. Janes' classification process, as also reported in the court's decision, estimated that about two-thirds of those were over-blocks. But since the sampling process drew from a set deliberately designed to have a very high concentration of OK items, it should be expected that a large percentage of the blocked items would also be OK. An even more fundamental problem occurred in studies presented by Hunter [1] and Lemmons [1, 10] that employed separate samples of OK and bad sites. Any estimate of the blocked-site overblock rate from such tests is arbitrary: selecting a larger or smaller sample of OK sites, while holding everything else constant, would yield different estimates of the blocked-site overblock rate.

If a study selects only blocked items for a test set, it cannot calculate the OK-sites overblock rate. To do that, one would need additional information about the proportion of

blocked to unblocked sites and the proportion of unblocked sites that were OK. For example, Edelman tested more than 500,000 URLs in order to select the 6,777 blocked items. If, as seems likely, the vast majority of the 500,000+ unblocked sites were acceptable, then the OK-sites overblock rate may have been under 1%. However, one can not be sure since the study was designed only to identify blocking errors, not their frequency among all OK sites.

Now consider the rate of under-blocking. One measure, which we call the *bad-sites underblock rate*, is the percentage of all unacceptable sites that were not blocked. This measure is related to recall in information science and sensitivity in medical research. It is the number that a school or library or parent or judge should consider when deciding whether blocking software is effective at preventing children from accessing pornography or other undesirable materials.

Another measure, the *unblocked sites underblock rate*, is the percentage of all unblocked sites that should have been blocked. This measure could be useful in determining whether an honor code is needed in addition to any installation of filters. For example, if this error rate is high, then the fact that a site was not blocked does not necessarily mean that it is non-pornographic, and it might be necessary to inform students that they are still responsible for not visiting porn sites even if the filters do not block their access. Again, the two measures of the under-blocking rate are independent: one may be high without the other being high. In Tables 1 and 2, the unblocked sites underblock rates are both 50%, but the bad-sites underblock rates are 1% and 99% respectively.

III. Conclusion

There have been numerous studies that report the over- and under-blocking rates of filtering software products. The methodology of such studies has improved substantially in recent years, but significant concerns still remain. Figure 3 summarizes desirable methods.

There is no easy answer to the question of how to best protect children from "inappropriate" material on the Internet [9], or even whether any protection is needed. Certainly filtering software is not a silver bullet. There are other approaches available, including student education, privacy screens, honor codes, and adult monitoring. However, the amount of attention and public concern about whether filters are helpful or harmful suggests an ongoing need for careful empirical investigation. Objective and methodologically sound research must inform the debate.

Values, however, will still have the final word. How much over-blocking or under-blocking is too much? When we reported finding of error rates in blocking health information [7], few questioned our methods or findings, but both supporters and opponents of filtering claimed that the results supported their positions. People simply differ in their assessments of the benefits of blocking bad sites and the costs of blocking OK sites. Methodologically sound research is needed to redirect attention away from meaningless debates comparing misleading study results toward meaningful debates about values.

Figure 3. Methodology Checklist

Test Set Selection

- Objective and repeatable process
- Independent of filters
- Large enough set to give statistical power

Blocking Test

- Redirects handled properly
- Updated blocking lists
- Multiple configurations tested

Classification

- Face validity
- Reliability
  - Criteria documented sufficiently to allow repetition
  - Inter-rater reliability reported
- Construct and criterion validity
- Sites cached at time of blocking tests

Error Rate Reporting

- OK-site overblock rate
- Blocked-site overblock rate
- Bad-site underblock rate
- Unblocked-site underblock rate

References

1. *American Library Association, Inc., v. United States*. 2002, E.D. PA.
2. Brunessaux, S., O. Isidoro, S. Kahl, G. Ferlias, and A.L.R. Soares, *Report on Currently Available COTS Filtering Tools*. 2001, MATRA Systemes & Information; Matra Global Netservices; Sail Labs; Hypertech; Red Educativa. Available online at http://www.net-protect.org/en/results3.htm.
3. Greenfield, P., P. Rickwood, and H.C. Tran, *Effectiveness of Internet Filtering Software Products*. 2001, CSIRO Mathematical and Information Sciences. Available online at www.aba.gov.au/internet/research/filtering/
4. Kleiner, A. and L. Lewis, *Internet Access in U.S. Public Schools and Classrooms: 1994-2002*. 2003, U. S. Department of Education, National Center for Education Statistics: Washington, DC. NCES 2004-011. Available online at http://nces.ed.gov/pubs2004/2004011.pdf.
5. Oder, N., *The New Wariness.* Library Journal, 2002. **127**(1): p. 55-57.
6. Pedhazur, E.J. and L.P. Schmelkin, *Measurement, Design, and Analysis : an Integrated Approach*. 1991, Hillsdale, NJ: Lawrence Erlbaum and Associates. 819.
7. Richardson, C., P. Resnick, D. Hansen, and V. Rideout, *Does Pornography-Blocking Software Block Access to Health Information on the Internet?* Journal of the American Medical Association, 2002. **288**(22).
8. Rideout, V., *Generation Rx.com: How Young People Use the Internet for Health Information*. 2001, Henry J. Kaiser Family Foundation: Menlo Park, California.
9. Thornburgh, R. and H. Lin, eds. *Youth, Pornography, and the Internet*. 2002, National Academy Press: Washington, DC.
10. U.S Department of Justice, *Web Content Filtering Software Comparison*. 2001, eTesting Labs: Morrisville, NC. Available online at http://www.veritest.com/clients/reports/usdoj/usdoj.pdf.