Web Page Classification: An Exploratory Study of the Usage of Internet Content Rating Systems

Michael G. Noll
Luxembourg International Advanced Studies in Information Technologies (LIASIT)
209, route d'Esch
1471 Luxembourg City, Luxembourg
noll@liasit.lu

Christoph Meinel
Hasso-Plattner-Institut für Softwaresystemtechnik,
University of Potsdam
14440 Potsdam, Germany
meinel@hpi.uni-potsdam.de

ABSTRACT

The greatest use of the Internet and new online technologies today is for constructive purposes. However, the use of the same technologies to spread illegal and objectionable content has been increasing dramatically in the last years. Internet users have begun to protect themselves and their wards by using so-called web content filters, which allow access to legitimate content and block access to objectionable, illegal, and otherwise harmful content. Next to active filtering technologies, which use heuristics, machine learning and similar techniques from the area of text and image classification to analyze web pages, there is the complementary category of passive content filters, which rely on (mostly voluntary) content rating systems to classify web pages. In the last years, content rating systems have received increased public attention and support, for example by initiatives such as the European Commission's Safer Internet campaign. In this paper, we study the usage of Internet content rating systems in the context of web page classification, and in particular the filtering of pornographic web content. Based on 8,000,000 anonymized Internet requests collected over a 1-month period, we have tested more than 150,000 websites for the presence of content rating information, so-called content labels; a random subset of 5,000 websites has been manually classified and used as the fundament for an evaluation of the classification performance of rating-dependent content filters for pornographic material in a real-world scenario. We show that the usage of Internet content rating systems is at best marginal and as a result of this, that the classification performance of rating-dependent content filters is inadequate and their application not yet recommended in practice.

1. INTRODUCTION

A recent web server survey[1] counted more than 63 million active websites and an average increase of 1.2 million sites per month for 2005. The greatest use of the Internet and new online technologies today is for constructive purposes; however, the use of the same technologies to spread illegal and objectionable content has been increasing dramatically in the last years. Internet users have begun to protect themselves and their wards by using so-called web content filters, which allow access to legitimate content and disallow access to objectionable, illegal, pornographic, and otherwise problematic content; for example, parents can use filtering software such as NetNanny or CyberSitter to safeguard their children from harmful websites. The problem of unwanted e-mails (unsolicited commercial e-mails, spam) has received increased public attention, and appropriate tools for filtering spam e-mails have been steadily integrated into the Internet's communication infrastructures and end-user applications. On the other hand, the area of web content filtering is still in its infancy. To improve this situation, various international, governmental and public initiatives have started campaigns such as the European Commission's Safer Internet programme to increase the public awareness of objectionable Internet content and support the development of technologies and frameworks to tackle harmful material on the World Wide Web.

One prominent approach is the usage of rating systems for Internet content, similar to rating systems such as MPAA's [2] for movies or ESRB's[3] for computer software and games. Like many of these, most of the existing Internet content rating systems are legally voluntary. Interested content providers can manually classify their content with a common description framework and add the rating information in the form of digital content labels to their websites. Internet users can then use filtering software to allow or disallow access to websites based on this meta information. Obviously, the availability of such content labels makes the filtering task *per se* rather trivial and theoretically more reliable than heuristic methods for content classification. For the rest of this paper, we will use the term "rating-dependent content filter" for filtering software, which relies only on this rating information to make filtering decisions.

Rating systems for Internet content sound promising on paper. But the viability and the success of content rating systems depend heavily on the actual usage of these systems

[1] Netcraft May 2005 Survey, http://www.netcraft.com/

[2] Motion Picture Association of America, http://www.mpaa.org/

[3] Entertainment Software Rating Board, http://www.esrb.org/

by the involved parties, in particular those responsible for providing rating information. To the best of our knowledge, the work in this paper is the first study analyzing the availability and trustworthiness of content rating information in the Internet. We show that the usage of Internet content rating systems is at best marginal and, at the example of using rating information to filter pornographic websites, that the resulting classification performance of rating-dependent content filters is inadequate and their application not recommended in practice.

2. RELATED WORK

The discussion about Internet content filtering has always been accompanied by censorship and privacy concerns. Recent studies [1] have shown that filtering technologies are one of the tools used by governments to restrict access to "inappropriate" Internet content. On the other hand, end users themselves have expressed their need for filtering technologies; for example, parents request better technical tools for protecting their children in the Internet [6], in particular for filtering pornography [5]. Ho and Lui analyzed the factors affecting Internet content filter acceptance [3] such as perceived usefulness in this context.

In an attempt to promote self-regulation of Internet content, rating systems have been introduced to help users control which content they want and do not want to see in the World Wide Web. Some related research work has been done on discussing the benefits and drawbacks of content rating systems in general [4], [8], in which voluntary rating systems [2] have been favored by most of the authors. To the best of our knowledge, the work in this paper is the first study analyzing the availability and trustworthiness of content rating information in the Internet.

3. INTERNET CONTENT RATING SYSTEMS

3.1 Overview

Internet content rating systems define special metadata to describe web content, so-called *content labels*. The creation of this metadata is generally performed on a voluntary basis by the content providers themselves, who will also technically integrate the rating information into their websites. Another though less common scenario involves third parties in the role of the content rating institution, who will classify content on behalf of others and provide this rating information on request.

Most of the existing Internet content rating systems are based on PICS, the Platform for Internet Content Selection[4]. PICS enables metadata to be associated with Internet content and promotes voluntary self-rating of online material [7]. It was originally designed to help parents and teachers control what children access on the Internet, and it is a platform on which other rating services and filtering software have been built.

The most prominent content rating system in the Internet today is developed and maintained by the Internet Content Rating Association (ICRA)[5], an independent non-profit organization established in 1999 by a group of international Internet companies and associations[6]. ICRA has been supported by the European Commission's Safer Internet Action Plan[7] and has participated in several EU funded projects in the fields of Internet security with a focus on content filtering. ICRA's current rating system is based on PICS but a successor using RDF (Resource Description Framework) is under development. The cornerstone of the rating system is the ICRA vocabulary[8], which defines a set of descriptors[9] used to classify online content. The vocabulary covers nudity and sexual content, violence, language, chat facilities, and other topics such as gambling, drugs, and alcohol. A selection of ICRA descriptors is listed in Table 1. In this paper, we focus on the ICRA content rating system for our studies.

Descriptor	Meaning
na 1	Erections and female genitals in detail
nd 1	Female breasts
ng 1	Obscured or implied sexual acts
nr 1	Appears in an artistic context and is suitable for young children
va 1	Sexual violence/rape
ve 1	Killing of human beings
vk 1	Deliberate damage to objects
lb 1	Crude words or profanity
oc 1	Promotion of drug use

Table 1: Selected ICRA content descriptors

3.2 Rating and filtering content

To rate material under their control, content providers use an online web form provided by the ICRA and check which of the (currently 45) elements in the ICRA vocabulary are present or absent from their websites. At the end of this process, the ICRA content label is automatically generated and can be integrated into the content providers' websites. The following label could be used to rate the content available at the LIASIT website, http://www.liasit.lu, and would be put into the <HEAD> section of every LIASIT web page for which it is valid.

<meta http-equiv="pics-label" content='(pics-1.1 "http://www.icra.org/ratingsv02.html" 1 gen true for "http://www.liasit.lu/" r (nz 1 vz 1 lz 1 oz 1 cz 1))'>

The fictitious label describes the content of LIASIT's website as rather innocuous:

- No elements listed in the category "Nudity and sexual material"
- No elements listed in the category "Violence"
- No elements listed in the category "Language"

[4]See http://www.w3.org/PICS/.
[5]See http://www.icra.org/.
[6]In the same year, ICRA superseded the older RSAC rating system.
[7]See http://europa.eu.int/information_society/activities/sip/.
[8]See http://www.icra.org/vocabulary/
[9]See http://www.icra.org/_en/faq/decode/.

- No elements listed in the category "Chat"
- No elements listed in the category "Other topics"

In our example, we generated the most common kind of label, whose scope is valid for the whole website ("gen true", an abbrevation of "generic true") and not only for specific pages. After this procedure, the rating of the LIASIT website would be completed and we could use the ICRA label tester application to verify the (technical) correctness of the content label.

By the use of this rating information, content filtering software can allow or block access to labeled websites based on the user's preferences via a simple matching process. Similar to the labeling process performed by the content providers, users employ the rating vocabulary to specify which types of content are deemed appropriate or inappropriate for them. If the parents of a 10-year old girl wanted to protect her from online pornography, they might decide to configure a rating-dependent content filter in a such a way that it would block access to any website with content in the "Nudity and sexual material" category.

3.3 Dealing with unrated content

An ongoing point of discussion is about how to deal with unrated content and which type of websites rating systems should focus on. The first and intuitive approach is that mainly unsuitable websites need to rate their content because the general goal of rating systems is to protect users from unwanted material. On the other hand, no regulatory jurisdiction can impose a rating system on content outside of its control, e.g. material from another country[10], and criminal content providers are unlikely to care for content rating at all [2]. The second approach therefore argues that it is rather the legitimate websites that need to use rating systems in order to express their "innocence". While the first alternative supports a policy to allow access to unrated content, the second implies the policy to deny access to unrated content since it cannot be trusted. In this paper, we have studied the consequences of both cases to deal with unrated content with regard to classification performance.

4. EXPERIMENTS

4.1 Data sets

The first task is to build a suitable test corpus of websites; these websites will then be checked for the existance and the validity of content labels. There are basically two ways to collect the raw data needed for the test corpus: the first possibility is to randomly select websites from so-called web directories such as the Open Directory Project[11]; the second possibility is to collect real-world data of users' WWW access, for instance by the means of WWW proxy log files, which can be used to compile a list of requested websites.

In the context of Internet content labels, the second alternative is more appropriate. The biggest advantage is the increased "coverage" of websites; websites with objectionable content are not likely to be listed in public web directories (especially true for illegal websites) but should definitely be included in the test corpus. Collection of real-world data also results in a better reflection of actual user behaviour and increases the chance to catch newly created websites, which are not yet listed in a web directory.

It is mandatory to use a sufficiently large and diverse user base to be able to generalize the results of such a test corpus and ensure its representativeness. The data sets used in the context of this paper are based on an anonymized collection of more than 8 million WWW requests of several thousands of users, collected over a 1-month period. The list of websites has been retrieved from this raw data to create the initial version of the test corpus. However it is recommended to verify the "validity" of the websites in this initial corpus, i.e. whether the websites actually exist; for instance, a website might have been shutdown during the meantime, or users might have entered incorrect URLs in their web browser application. The total number of websites in the test corpus after such a verification check is 152,617; the number of unique top level domains[12] in the test corpus is 151.

For analyzing the usage of content labels in this paper, two test corpuses have been built (see Table 2). The first corpus, TOTAL, contains the total of 152,617 websites. The second corpus, RANDOM5000, is a random subset of the TOTAL corpus; it contains 5,000 websites, which have been manually classified into the categories "pornography" (14.2% of the corpus) and "not pornography" (85.8% of the corpus). This allows a more specific analysis of the usage of content rating systems for each of the two categories.

Corpus	number of websites
TOTAL	152,617
RANDOM5000	5,000
of which are pornographic	708
of which are not pornographic	4,292

Table 2: Overview of data sets (corpuses)

For making the manual classification of the websites in RANDOM5000 as objective and comparable with other studies as possible, we used the descriptors found in the ICRA vocabulary and classified any website as being pornographic, which had content in the "Nudity and sexual material" category and which did not contain the "inoculating" descriptors *nr1*, *ns1*, and *nt1*. The latter are used to describe content in the "Nudity and sexual material" category, which "appears in an artistic, educational, or medical context and is suitable for young children"[13].

[10] Just recently, the British media regulator Ofcom has considered industry-wide classification systems to help consumers better understand the suitability from TV shows to online videos and music downloads. However, Ofcom has been warned that classifying non-UK material on the Internet could prove problematic as it would not be covered by the same rules and legal framework. In addition, commercial UK broadcasters have reportedly claimed that such a system could dilute their brands, so that widespread support for Ofcom's initiative has yet to be achieved.

[11] See http://www.dmoz.org.

[12] For example, "com" or "org".

[13] ICRA provides a more detailed explanation of these descriptors: "The first part of the statement refers to the *intention*. Classical painting and sculpture can be assumed to be *intended as artistic*. Material designed to teach children about sex, would qualify as *intended as educational*.

It has to be noted that only the start pages of the websites in the corpuses were tested for labels. Though it is possible that other web pages of the same website do contain labels while the start page does not, this case is rather unlikely in practice. Not only because the start page is arguably the most prominent web page of a website, but also because ICRA content labels can be conveniently generated in such a way that their scope extends to any web pages hierarchically below the current (labeled) one.

4.2 Setup

An automated software tool has been developed by the authors to facilitate content label tests for websites. In order to ensure the correctness of our experiments, we configured the tool to query the official ICRA label tester web application[14] in "strict rules" mode for each website in the corpuses and return the official ICRA label test result.

Basically, the label test consists of three sub-tests: first, it tests a website for the presence of a content label (label presence); second, it verifies the syntactical correctness of the labels (label syntax); third, it verifies whether the complete website is labeled including any elements such as hyperlinked images (label coverage). A label tester result of "red" means that either no label has been found at all or only labels with errors[15] were present; "yellow" indicates a partially but not fully labeled website, i.e. although the website carries a label, some elements such as images or banners are not labeled or covered by the existing label. A "green" result is returned for a fully rated website with a syntactically correct label. In addition to these three results, we have included a fourth result "error" which indicates the failure of the label test[16]. It is important to note that the ICRA rating system counts "yellow" websites as unrated, which means that a filter relying on content labels would deny access to a web page if it was set to block unrated sites.

5. RESULTS

5.1 Availability

A total of 152,617 domains have been tested for the presence of ICRA content labels. The results for the availabilty of rating information are shown in separate tables for each of the corpuses.

Table 3 shows the test results for the TOTAL corpus. Only a marginal fraction of 0.6% of the websites is fully labeled; even when we include partially labeled websites in the calculation, the percentage of rated websites is only 1.5%. In other words, a rating-dependent web content filter would not be able to make any meaningful classification decisions for 98 out of 100 websites. Of course, content filters can be configured to either generally allow or deny access to unrated websites to overcome this problem of indecisiveness but with questionable outcomes (see section 5.3).

Test result	number of websites	percentage
Red	139,988	91.7%
Yellow	1,325	0.9%
Green	857	0.6%
Error	10,447	6.8%
Total	152,617	100.0%

Table 3: Label results for TOTAL

Test result	number of websites	percentage
Red	4,603	92.1%
Yellow	48	1.0%
Green	27	0.5%
Error	322	6.4%
Total	5,000	100.0%

Table 4: Label results for RANDOM5000

The RANDOM5000 corpus with its manually classified websites can be used to derive more specific results about the usage of Internet content rating systems. The overall statistics of this corpus (Table 4) reflect the numbers of the TOTAL corpus, showing again only a small fraction of rated websites. The category-specific statistics of the pornographic and non-pornographic websites in the RANDOM5000 corpus are shown in Table 5 and Table 6, respectively.

Test result	number of websites	percentage
Red	629	88.8%
Yellow	32	4.5%
Green	16	2.3%
Error	31	4.4%
Total	708	100.0%

Table 5: Label results for pornographic websites

In relative comparison, websites in the pornography category are much more likely to be labeled than non-pornographic websites, regardless whether we count both yellow and green (6.8% : 0.7%) or green only (2.3% : 0.3%) test results. A possible explanation could be that providers of pornograhpic content are more aware of Internet content rating systems than other providers and are also willing to make use of them, a rather promising discovery. Still, the overall usage of rating systems in the Internet as shown is only very marginal today.

5.2 Trustworthiness

A content rating system must make sure that its labels accurately reflect the actual content they describe in order to engender and maintain public trust in the rating system itself. The operators of rating systems such as ICRA therefore

However, the context descriptor also requires a (subjective) pledge that the material is *suitable for young children*. Material of an explicit violent or sexual nature, even if intended as artistic/medical/educational, may *not be suitable for young children*. This additional statement encourages caution when claiming 'redeeming context' [...] The ICRA international reference group strongly suggested that the categories should reflect a parents concern for young children. For many members of the expert group, *young* was perceived as under the age of 12."

[14]The label tester web application can be found at http://www.icra.org/label/tester/.

[15]For example, a typographic error in a URL definition of a label.

[16]For example, the label test for a website can fail because of network connection problems to the corresponding web server at the time of the test.

Test result	number of websites	percentage
Red	3,974	92.6%
Yellow	16	0.4%
Green	11	0.3%
Error	291	6.8%
Total	4,292	100.0%

Table 6: Label results for non-pornographic websites

claim to verify the correctness of their content labels in a periodical and automated way[17]. In our study, we compared the content labels of the fully rated "green" websites in the RANDOM5000 corpus (Table 4) for both the "pornography" and "not pornography" categories with the visible content on these websites to measure the trustworthiness of content rating systems.

A website is correctly labeled if and only if its label information matches the visible content of the website, which implies that the website needs to be fully labeled. If l_c and l_f are the numbers of correctly labeled and incorrectly labeled websites, respectively, we have defined trustworthiness as

$$trustworthiness = \frac{l_c}{l_c + l_f} \tag{1}$$

In contrast to the quantitative measurement of availabilty (see section 5.1), trustworthiness can be interpreted as a qualitative measurement for rating information and, indirectly, for its corresponding rating system.

We found discrepancies in 18.5% of the cases, i.e. label and content did not match, resulting in a trustworthiness of 81.5%. The majority of the discrepancies was caused by pornographic websites, whose label descriptions were too lax for the available content; for example, a label included only the descriptor for "Female Breasts" while the content showed genitals in detail. Surprisingly, there was even a case of a clearly non-pornographic website, which incorrectly contained a label describing pornographic content.

Our results suggest that it is not advisable to blindy trust in available content labels. The basic theoretical assumption that every fully rated ("green") website can be correctly classified is thus not generally true in practice, which further lowers the potential classification performance of rating-dependent filters.

5.3 Classification performance

In the following sections, we want to assess the performance of rating-dependent content filters for the porn classification task. Rating-dependent content filters can only make a meaningful classification decision for a given website if its content is fully and correctly rated. However, we have seen that only a small fraction of Internet content has been rated and contains valid labels. Before we can evaluate the classification performance of rating-dependent content filters, we need to answer the question how to deal with unrated content. There are two options available, both of which will be discussed here: to set a web content filter to either generally allow or deny access to unrated material. For the following calculations, only fully labeled "green" websites count as rated; the remaining websites including those with an erroneous label test result count as unrated.

[17]See http://www.icra.org/trust/ for more information about how ICRA ensures the correctness of its content labels.

5.3.1 Performance criteria

Classification performance is usually measured in terms of the classic Information Retrieval notions of *recall* and *precision*, adapted to the case of text classification. For the following sections, *FP* (false positives) is the number of legitimate, non-pornographic websites incorrectly classified as pornographic; *TN* (true negatives), *TP* (true positives), and *FN* (false negatives) are defined accordingly (see Table 7). A rating-dependent content filter will classify a given website as pornographic if and only if 1) the website is fully labeled and its label describes the visible content as pornographic, or 2) the website is not fully labeled and the filter is set to deny access to unrated websites.

		Expert judgement	
		porn	**not porn**
Classifier judgement	**porn**	TP	FP
	not porn	FN	TN

Table 7: Contingency table for website classification decisions

Based on the notations above, we can define *recall*, *precision* and F_1 for pornographic websites as:

$$recall = \frac{TP}{TP + FN} \tag{2}$$

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$F_1 = \frac{2 * recall * precision}{recall + precision} \tag{4}$$

The F_1 score is the (equally weighted) harmonic average of *recall* and *precision*. In addition, we define the $false\ positive\ rate$ and $false\ negative\ rate$ as

$$false\ positive\ rate = \frac{FP}{TN + FP} \tag{5}$$

$$false\ negative\ rate = \frac{FN}{TP + FN} \tag{6}$$

5.3.2 Performance results

The first case is a filter, which is set to generally allow access to unrated content. This alternative will never create

false positives for unrated websites (since access is always allowed) but on the other hand it will cause a rather high amount of false negatives. Column 1 in Table 8 shows Recall and Precision for the first case. Precision is at 100.0% because the filter will *per definitionem* only block websites with a valid pornographic label; the drawback is that the recall of 1.7% for this filter is so poor that it is almost never able to identify a de-facto pornographic website as such. Because only a small fraction of websites is actually rated, the first alternative is only marginally better than not using a filter at all.

The second case is a filter, which generally blocks access to unrated content. In contrast to the former case, this filter will never create false negatives for unrated websites (since access is always blocked) but on the other hand it will cause a high amount of false positives. This translates to a perfect recall of 100.0%, which means that access to every single pornographic website is being blocked; but the drawback is that almost all legitimate websites fall prey to the filter, too.

Unrated content will be	**allowed**	**blocked**
Recall	1.7%	100.0%
Precision	100.0%	18.9%
F_1	3.3%	31.8%
False positive rate	0.0%	99.8%
False negative rate	98.3%	0.0%

Table 8: Classification performance of rating-dependent content filters (depending on the handling of unrated content)

Regardless of which alternative is chosen, the result is not very satisfying. In the first case, the filter is unable to fulfill its purpose of eliminating objectionable content because it misses more than 98% of pornographic websites; it is only slightly better than not using a web content filter at all, i.e. allowing access to any website. In the second case, the filter blocks access to more than 99% of legitimate websites; its effect is almost the same as cutting off access to the Internet, i.e. blocking access to any website. Of course, the main reason for the poor classification performance is the low usage of content rating systems in the Internet.

6. CONCLUSIONS

In this paper, we have studied the usage of Internet content rating systems in the context of web page classification, in particular the filtering of pornographic web content, by analyzing more than 150,000 websites. We have shown that the usage of content rating systems today is only marginal in practice; a very small fraction of the large set of analyzed websites contained rating information, and we could assert correct rating information for even less. As a result of this, the classification performance of rating-dependent content filters is very poor; based on whether content filters are set to allow or disallow access to unrated content, the performance is only slightly better than not using a filter at all or blocking access to any website, respectively. Based on our findings, the fictious pair of parents from section 3.2 would be advised not to rely on rating-dependent content filters to protect their little daughter, at least not without another tier of security[18]. It has to be noted that despite our findings Internet content rating systems such as ICRA are a promising approach; but their popularity is as yet not high enough to make them self-sufficient and viable in practice.

We also stress that operators of content rating systems must make sure that rating information correctly reflects the labeled content; we have shown that this is not always the case in practice. It is mandatory that users can trust in the correctness of content rating information (if it is present), otherwise they might reject the rating system completely.

7. REFERENCES

[1] Internet filtering in china in 2004-2005. Technical report, OpenNet Initiative, 2005.

[2] J. M. Balkin, B. S. Noveck, and K. Roosevelt. Filtering the internet: A best practices model. Available at http://islandia.law.yale.edu/isp/filtering/overview.html. Information and Society project, Yale Law School.

[3] S. Y. Ho and S. M. Lui. Exploring the factors affecting internet content filters acceptance. *ACM SIGecom Exchanges*, 4:29–36, 2003.

[4] C. D. Hunter. Negotiating the global internet rating and filtering system: Opposing views of the bertelsmann foundation's self-regulation of internet content proposal. In *Proceedings of the tenth conference on Computers, freedom and privacy*, pages 235–238, 2000.

[5] P. Johnson. Pornography drives technology: Why not to censor the internet. *Federal Communications Law Journal*, 49:217, 1996.

[6] S. Livingstone and M. Bober. Final report of key project findings. Technical report, UK Children Go Online, 2005.

[7] J. Varghese, R. Krishnan, Y. U. Ryu, R. Chandrasekaran, and S. Hong. Filtering objectionable internet content. In *Proceedings of the 20th international conference on Information Systems*, pages 274–278, 1999.

[8] J. Weinberg. Rating the net. *Hastings Communications and Entertainment Law Journal*, 19:453, 1999.

[18] Of course, parental education should never solely depend on technical solutions.