OpenChoice: A Platform for Web Content Classification & Filtering

Don Turnbull & Miles Efron
School of Information, University of Texas at Austin
1 University Station, D7000 Austin, TX 78712 USA
donturn@ischool.utexas.edu & miles@ischool.utexas.edu
http://www.ischool.utexas.edu/~choice

Many organizations would like to be able provide a filter for Web information be it blocking advertising, pop-ups or for providing kid-safe access on public computers. The current state of commercial filters provides little, if any, control over the type of information each organization may need to filter. The ***OpenChoice*** system currently in development, will provide an open source filtering system to allow organizations to configure and tune for their own Web information filtering requirements. Using webs of trust and open source statistical modeling software, *OpenChoice* is intended to further the open source community's efforts to uncouple compliance with CIPA from contracting with commercial software vendors. *OpenChoice* attempts to re-imagine content filtering as a form of collection development to be practiced by system administrators and information professionals working in diverse environments.

Impact

This research and demonstration project will address the issue of filtering by creating, testing, and evaluating an open source Internet content filter. While no response to filtering will satisfy all parties, this study will examine the open source model of software development as a possible "third way" of addressing the debate. Developed by volunteers and distributed with minimal intellectual property restrictions, open source software has proven its value[i] in the arenas of operating systems (e.g. Linux[ii]), server software (Apache[iii]), and database management technology (MySQL[iv]).

Besides its wide applicability, though, *OpenChoice* will serve the field by obviating two of the most fundamental arguments against filtering software: financial cost and intellectual property constraints. While commercial filtering software is expensive to purchase and maintain (due to ongoing subscription fees)[v], as an open source project *OpenChoice* will be available free of charge. *OpenChoice* will assuage the concern among many information professionals that commercial filtering products rely on proprietary databases and algorithms that are considered trade secrets.[vi] Hence, many who are forced to use these commercial filters bristle at their compromised role as information stewards, as they cannot compare or tune filters'. The project will also develop a community portal to allow any participants to exchange expertise about crafting appropriate "block lists." *OpenChoice* will rely on state of the art machine learning technologies to assess the suitability of requested content. This automatic classification module of the system will be based on open source machine learning tools[vii], and will use no proprietary software. *OpenChoice* outlines a model program whose design can be expanded and/or adapted to include a wide range of information settings interested in the issue of Internet filtering or other issues that might benefit from the development of open source software. As a model of cooperation it illustrates how practitioners and researchers can collaborate to build an online community to bring workable solutions to a field-based challenge. Our plan also includes the development of Web-based training modules.

Design

The Content Filtering Application

OpenChoice will run as a proxy server. When patrons request a resource from the Internet on a library terminal, their request passes through the proxy before being fetched from the Web. The filtering application tests the requested URL against its internal list of banned resources. If the item is not present in the set of banned resources, the transaction proceeds; otherwise, the user's request is denied and their browser is redirected to a Web page that invites them to request that a librarian review the decision to block.

Pending further analysis, the investigators plan to build the *OpenChoice* proxy server from the pre-existing open source proxy called SquidGuard[viii]. SquidGuard was chosen because it is distributed with almost no restrictions on re-development, and because it is built on top of established, open standards. In particular, SquidGuard is closely related to the Apache Web server, a leading project in the open source movement that sees vigorous ongoing development. A filter using SquidGuard as a proxy to Apache will thus rely not only on open transmission standards (HTTP, TCP/IP), but will also be based on software with strong user and development bases.

The primary change that will be made to SquidGuard in fashioning the *OpenChoice* filter is the application's method of constructing its URL blacklist. SquidGuard currently operates by use of a "dumb robot." That is, it employs a web crawler that adds URLs to the block-list based on extremely rudimentary criteria.

To improve the system's accuracy, *OpenChoice* will submit the resultant blacklist to two forms of vetting to alleviate the problem of over-blocking. At the level of the filtering application improvement will come from the application of machine learning-based document analysis. After acquiring a new, putatively harmful URL, the application will estimate the likelihood (using state-of-the-art filtering technology) that the URL is actually harmful. Newly acquired URLs will thus be ranked according to the system's confidence that they are in fact harmful. This ranked list will then be vetted by a community of information professionals via the Web-based portal.

The Web-Based Portal

OpenChoice is predicated on the idea that a judicious combination of automated filtering and human judgment will lead to a superior filter and a feeling of professional investment on the part of librarians. To foster these goals, volunteer librarians, system administrators, users and computer scientists will contribute to the performance of *OpenChoice* by voting on the appropriateness of the system's newly-acquired questionable URLs. The goal of the portal is to encourage everyone to take an active role in crafting *OpenChoice*'s configuration. Users of the portal will participate by creating personal system accounts. When a user logs into the system, he or she will see a ranked list of those resources most in need of human review (i.e. those that the learning algorithm is least confident about). The user will then "vote" on as many of these URLs as he or she desires. Once the votes on a particular URL reach a critical mass of consensus, that URL is added to the canonical *OpenChoice* blacklist. All additions are subject to future review at the suggestion of any community member.

A potential objection to this vetting process is that the blacklist's quality might be open to sabotage from the volunteer editing process. To obviate this problem, the system will rely on a trust model such as those used by contributor-run digital libraries[ix]. Under models of trust, each volunteer's contributions are implicitly judged by the community as a whole[x]. New editors enter the system with very little "clout"; their votes are considered provisional, pending review by established editors. As a user participates in the system over time, however, his or her clout increases if his or her votes are frequently in agreement with the mainstream of *OpenChoice* volunteers. Such models have shown been to organize information effectively and fairly. Systems based on so-called webs of trust have two distinct advantages. First, they use the collective efforts of a user community to improve the community's control over information. Second, they provide an incentive for members of the community to participate in the system's improvement by allocating social capital to those community members who participate meaningfully. Figure 1 depicts the flow of actions and information within the *OpenChoice* system:

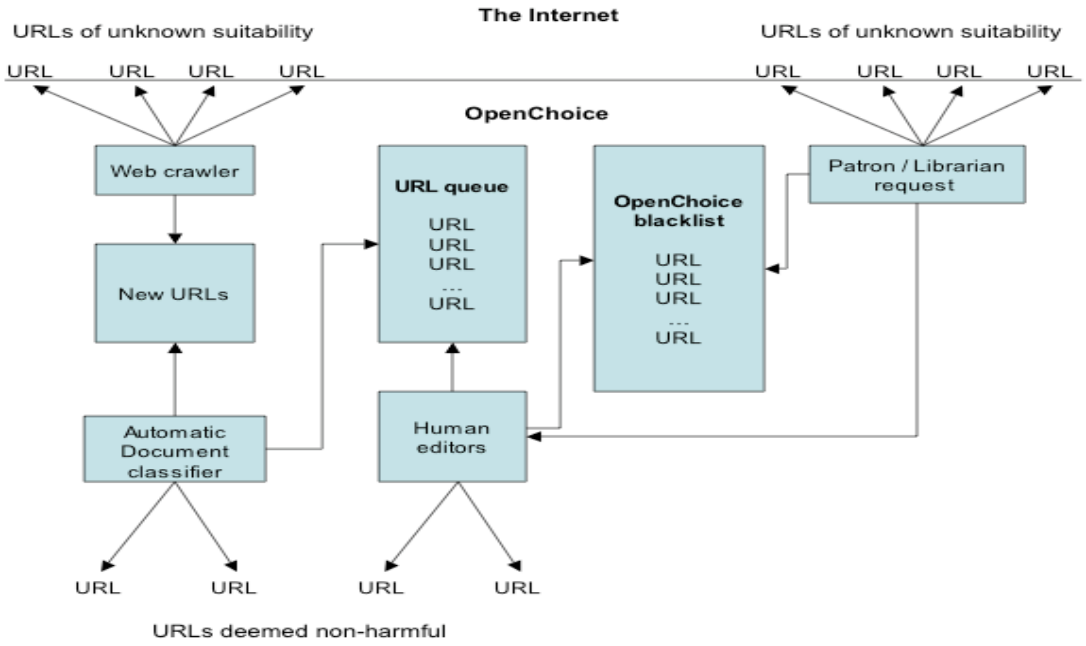

Figure 1: Information Flow within *OpenChoice*

Outcomes

The *OpenChoice* project will address this research question: Can open source models of software development produce and maintain an Web content filter that addresses the needs of practicing information professionals? Answering this question involves two high-level goals: 1) to develop an open source Internet content filter for libraries and 2) to develop software that allows a community to contribute to the development of the filter. This poster is intended as a first step in publicizing the project to garner a community of users, developers and volunteer editors. To determine the success of *OpenChoice*, multiple methods of evaluation, some of which are briefly described in the following table.

Goal	1. Developing the Filter	2. Developing a Community Portal
Measurable Units	See *Quality Control* section below	Community Size: Number of registered users in the *OpenChoice* portal
		Community Activity: Number of contributions to the blacklist
Data Sources	Existing filter blacklists and pre-classified data	Internally managed database of users
Observation Intervals	During development: as features are added to the filter	Daily
	After development: as community participation reaches benchmarks	
Target	Accuracy equal to non-collaborative free filters and commercial filters	Comm. Size: 200 members within 1 year of initial dissemination efforts[xi].
		Comm. Activity: A density of contribution sufficient to estimate their distribution

Table 1: Output-Based Measures for *OpenChoice* Project Goals

Of utmost concern is the quality of the filtering itself. Does the filter block inappropriate material? Does the filter allow users to view materials that the community accepts? While these issues will also surface in our outcome-based evaluation, the *OpenChoice* project will entail ongoing, rigorous research into the error rate of the developed software. This evaluation will follow the well-defined research paradigm in machine learning where the goal is to estimate the error rate of a classifier on the population of unseen data. To make this estimation, classified data (acceptable/unacceptable) are presented to the statistical model and the results are tabulated. In this manner, researchers gain an unbiased estimate of the classifier's error rate. Furthermore, this statistic is a simple proportion of accurate to inaccurate classifications. As such, traditional exact tests on the difference between proportions can be conducted to compare the performance of any two classifiers.

Evaluating the quality of *OpenChoice* will involve gathering and testing a large test collection of URLs. These will be drawn from publicly accessible repositories of pre-classified Web resources such as Yahoo! For instance, we will add to the test collection's list of unacceptable sites those URLs in Yahoo's *pornography* and *recreational drugs* sections, among others. Likewise collecting URLs from more innocuous portions of the taxonomy will develop a list of putatively acceptable resources. The goal will be to derive a test set of at least 2000 URLs in order to derive adequate statistical power. Once the test set is in place, the error rate of *OpenChoice's* classifier will be evaluated against the ground truth of the human-classified documents. Likewise, we will obtain estimates of the error rates of SquidGuard using the un-improved blacklist, as well as the most popular commercial filters. Finally, at each phase of evaluation, we will test the significance of the difference in error rates between each filtering product. The goal of this evaluation is to gauge the effect of *OpenChoice's* two mechanisms for improving a filter: 1) machine learning for identifying putatively misclassified items in the blacklist and 2) community involvement in improving the blacklist. We hypothesize that applying these measures to an open source filter will improve the filter's accuracy.

As with any open source project, the developers of *OpenChoice* will hope for contributions from the open source community to continue ongoing development of the filter's blacklist. Because the filter's underlying technologies (WEKA, Squid, MySQL) are themselves all open source projects, the costs of improving the service over time will bear mainly on the efforts of the maintainers and the community of developers and participating filtering community.

---

[i] Raymond, E. (1999). *The Cathedral and the Bazaar.* Sebastepol, CA: O'Reilly.
[ii] The Linux Kernel Archives. Retrieved 25 October 2004 from http://www.kernel.org
[iii] The Apache Project. Retrieved 25 October 2004 from http://www.apache.org
[iv] MySQL. Retrieved 25 October 2004 from http://www.mysql.com
[v] Ayre, L. B. (2001). Internet filtering options analysis: An interim report. InfoPeople Project. Retrieved 25 October 2004 from http://www.infopeople.org
[vi] Goldstein, A. (2002). Like a sieve: The Child Internet Protection Act and ineffective filters in libraries. *Fordham Intellectual Property, Media, and Entertainment Law Journal,* 12, 1187—1202.
[vii] WEKA: The Waikato Environment for Knowledge Acquisition. http://www.cs.waikato.ac.nz/ml/weka/ retrieved 25 October 2004. WEKA is an open source library of machine learning algorithms implemented in JAVA. The PI has conducted extensive research using WEKA's libraries for text mining and management.
[viii] http://www.squidguard.org retrieved December 5, 2004.
[ix] Jones, P. (2001). Open(source)ing the Doors for Contributor-run Digital Libraries. *CACM* 44(5), pp. 45-46.
[x] Rheingold, H. (2002). *Smart Mobs: the Next Social Revolution.* Perseus Books.
[xi] Roughly equal to the number of subscribers to the SquidGuard users mailing list.