Available online at www.sciencedirect.com

ScienceDirect

Decision Support Systems 42 (2006) 144–161

Decision Support Systems

www.elsevier.com/locate/dsw

A method for managing access to web pages: Filtering by Statistical Classification (FSC) applied to text

Jonathan P. Caulkins [a], Wenxuan Ding [b], George Duncan [a], Ramayya Krishnan [a,*], Eric Nyberg [c]

[a] *The Heinz School, Carnegie Mellon University, Pittsburgh, PA 15213, United States*
[b] *Department of Information and Decision Sciences, and of Computer Science, University of Illinois, Chicago, IL 60607, United States*
[c] *The School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, United States*

Available online 18 January 2005

Abstract

Various entities (e.g., parents, employers) that provide users (e.g., children, employees) access to web content wish to limit the content accessed through those computers. Available filtering methods are crude in that they too often block “acceptable” content while failing to block “unacceptable” content. This paper presents a general and flexible classification method based on statistical techniques applied to text material, that we call, Filtering by Statistical Classification (FSC). According to each individual entity’s expressed opinions about what content in a training data set is or is not acceptable, FSC constructs a customized model to represent each individual entity’s preferences. FSC then uses this customized model to examine new web content and to block unwanted content. The empirical results suggest that our method has greater predictive power than do a variety of existing approaches.

Keywords: Content-based filtering; Decision support; Statistical classification techniques

1. Introduction

Those who use electronic technology to access information often do not own, control, or otherwise have responsibility for the means of access, nor do they bear the full consequence of its use. For example, employees may use an employer’s computer, children may use the family’s computer, and the public may access the Internet through library computers. The responsible entity may have an obligation or a preference to regulate what information the user accesses through its facilitation. Prototypical applications are parents protecting their children from age-inappropriate content (e.g., glorification of drug use,

* Corresponding author. Tel.: +1 412 268 2174; fax: +1 412 268 7036. The order of authorship is alphabetical.

E-mail addresses: caulkins@andrew.cmu.edu (J.P. Caulkins), wxding@uic.edu (W. Ding), gd17@andrew.cmu.edu (G. Duncan), rk2x@andrew.cmu.edu (R. Krishnan), ehn@cs.cmu.edu (E. Nyberg).

0167-9236/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2004.11.015

instructions on creating bombs, etc.), employers concerned that users might create a hostile work environment for colleagues by accessing inappropriate material such as pornography, and employers preventing employees from spending work time on non-work related content (comics, stock quotes, sports scores, etc.). In the non-electronic context, such filtering is usually non-controversial. A good parent selects age-appropriate books for their children. Advised of a potential for legal liability, employers prevent employees from displaying pornographic magazines and posters in the workplace. Managers seek to keep employees working, not shirking by reading a newspaper's sports pages.

Our goal is to help these responsible entities–call them *guardians*–exercise their legitimate authority to control in the electronic domain what they already control in the physical domain. In an informational context, this means providing an effective filter that provides full and ready access to appropriate material, while blocking access to inappropriate material. This is not an easy task. The variety and extent of material available to the user are so vast, and changing so rapidly, that the guardian cannot possibly catalog and separate the objectionable material from the acceptable. The challenge is to block access to material that a guardian would deem objectionable, even when it is not feasible for a guardian to assess directly the suitability of each site. Conceptually, we address this problem as a classification task. Based on available information, should a site be displayed or withheld? Other categories of action are possible, such as the intermediate one of sending the content to the guardian for evaluation.

We use the specific domain of filtering pornographic content on the web to illustrate our method, because child access to Internet pornography is considered a serious problem [43] and because the inherently visual aspect of much pornography makes it a particularly challenging application (general information about family control of children's access to the Internet is available on the website, GetNetWise (http://www.getnetwise.org/) which as of August 2004, listed more than 65 different software tools for filtering sexual material.) Our method is applied to the text of the subject site, rather than its visual elements. Others are attempting to solve the specific problem of identifying pornographic visual content ([45], also, http://www-db.stanford.edu/pub/gio/2001/wipe-forum.ppt). These efforts are more complementary than competing for at least two reasons. First, they apply only to visual pornography, whereas text-based methods are equally suitable for screening pornographic stories, hate group tracts (for children), sports pages (for office workers), and other classes of material deemed inappropriate. Second, a combination of both approaches is likely to be more effective than either in isolation. Image-based methods may fail when the individuals are partially clothed, and the text-based methods may fail if there are few words on the page, but the chance of both methods failing is lower.

Current methods for filtering based on text include combinations of (1) blacklisting "bad" or whitelisting "good" sites, (2) blocking sites that include "bad" words, and (3) using a "rating" of the web site, whether given by the site creator or a third party, (e.g., the Platform for Internet Content Selection or PICS approach http://www.w3.org/PICS, [36]). These methods have limitations. In the dynamic environment of the web, Method (1) is inadequate because no list of "bad" or "good" sites can be current. Likewise, Method (3) cannot be comprehensive because only a small percentage of web sites can be rated. Furthermore, different people (or guardians) may have different opinions on whether a given web page should be blocked or not, and those differences may not be reflected fully in any simple scale. Method (2), keyword-based filters, breaks down because many "good" sites include "bad" words. Some filtering software uses elaborate context rules, for example, "breast" is bad, unless it appears as "breast cancer," but again, the appropriate context rules can vary by application and even by guardian. For example, Lee et al. [25] use neural networks to filter pornographic web pages. However it does not attempt to customize filtering to fit the value judgments of the guardian. Hence, there is a need for a flexible method that can adapt to each guardian's preferences.

Hence, we have developed a new method, that we call Filtering by Statistical Classification (FSC), to enable personalized control in managing access to web sites. The method uses statistical classification tools through an analysis of certain key features of the subject material in a training data set. The method is (1) *comprehensive*, in that it applies to *all* sites, (2) *customizable*, to reflect the values of a