Automatic Web Rating: Filtering Obscene Content on the Web

K. V. Chandrinos, Ion Androutsopoulos, G.Paliouras, and C. D. Spyropoulos

Institute of Informatics and Telecommunications
National Centre for Scientific Research "Demokritos"
153 10 Ag. Paraskevi, Athens, Greece
{kostel, ionandr, paliourg, costass}@iit.demokritos.gr

Abstract. We present a method to detect automatically pornographic content on the Web. Our method combines techniques from language engineering and image analysis within a machine-learning framework. Experimental results show that it achieves nearly perfect performance on a set of hard cases.

1 Introduction

Pornography on the Internet, although less abundant than certain news reports have claimed, is a reality.[1] To cater for the broader problems of content characterization, the World Wide Web Consortium has introduced the Platform for Internet Content Selection (PICS) [6], a mechanism that allows Web pages to be rated in many dimensions (e.g. violence, nudity, suicidal content). PICS can be used either by Web authors that want to label their sites with metadata describing their content, or by third-party rating authorities. Pornographic sites are covered by at least two rating schemes, one from ICRA [7] and one from SafeSurf [8]. Once a page has been rated under PICS, popular browsers can be configured to take this rating into account. It is clear, however, that many pornographic sites are not willing to adopt self-regulation; and client-side configurations of off-the-shelf browsers can be easily circumvented.[2] Evidence to the fact that Internet users may elect to block pornography for themselves or minors under their responsibility comes from a number of commercial solutions. These include proxy servers that check requests against a list of forbidden or allowable URLs, and client-based software that utilizes a combination of blacklists and shallow keyword-based analysis. The dynamic nature of the Web and the fact that it is extremely easy to set up or migrate a Web site to a new IP address makes the blacklist approach ineffective.[3] To counterpart this, current commercial solutions dispatch updates of their lists to their customers, with list lengths reaching up to a few hundred thousand URLs. Querying Altavista with the keyword "sex", however,

[1] See http://websearch.about.com/internet/websearch/library/myths for related statistics.

[2] Consult [4] for an evaluation of third-party and self-regulating rating schemes.

[3] An evaluation copy of a commercial product included a ~12,000 strong URL blacklist which contained only 36 out of the 500 pornographic URLs we easily summoned in a day.

J. Borbinha and T. Baker (Eds.): ECDL 2000, LNCS 1923, pp. 403–406, 2000.

returned roughly 9 million hits on average during May 2000. To alleviate list inefficiency, existing blocking software often includes keyword scanning of the URL and/or the title of the requested document. The keyword list is manually constructed to reflect frequently used terms in pornographic sites, and can often be augmented by the end-user. While real-time keyword scanning of the URL and title can improve underblocking by trapping sites not included in the blacklists, it introduces a serious amount of overblocking. For example, many Web sites maintaining content on *sexual* harassment or discrimination, *rape* prevention or even *oral* hygiene, are indiscriminately blocked under such a scheme.[4]

Our approach combines language engineering and image analysis within a machine-learning framework. It uses a probabilistic classifier trained on an appropriate corpus of Web pages, and employs both textual attributes and attributes derived from the results of an image processor. The latter estimates whether significant parts of the images on a Web page contain skin tones and are therefore suspected of depicting nude subjects. Our method is automatic in the sense that, once trained, our filter does not require Web pages to be rated manually.

2 Filtering Techniques

We represent each Web page as a vector $\vec{x} = \langle x_1, x_2, x_3, \ldots, x_n \rangle$, where $x_1, \ldots, x_n$ are the values of attributes $X_1, \ldots, X_n$. All attributes are binary, i.e. $x_i = 1$ if the page has the property represented by X_i, and $x_i = 0$ otherwise. $X_1, \ldots, X_n$ are selected from a pool of candidate attributes that includes both textual and image attributes. Textual candidate attributes correspond to words, i.e. each textual candidate attribute shows whether or not a particular word (eg. *adult*) is present on the page. There are currently only two image attributes: one showing whether or not the page contains at least one suspicious image (IMG1), and one showing whether or not the pages contains at least one non-suspicious image (IMG2). We use *mutual information* [12] to select the best attributes from the pool, and train a Naive Bayesian classifier [1] [3] to distinguish between vectors that correspond to obscene (pornographic) and non-obscene pages.

There are extensive reports in the literature on methods for skin detection, since this is a critical step towards face detection and recognition [10, 11]. These methods rely mostly on color-space transformations from RGB to YC_bC_r or HSV where the range of skin tones is better constrained. Such transformations, however, present a trade-off between accuracy and computational expense. The only work that looks into flesh tones for the identification of naked people is that by Forsyth and Fleck [2], which attempts to use color and geometrical heuristics to detect humans. The geometrical constraints bring down recall from 79% to a mere 43%, minimizing false positive responses to 4%. Since accuracy of human bodies contouring is not as critical for our application as speed, we developed a fast and robust estimator that indicates in a single-pass whether or not more than a certain percent of an image depicts skin tones utilizing information solely from the RGB space. In our configuration, the presence of an image that has been judged to be possibly pornographic cannot alone force the

[4] Blocked sites include Christian Bookselling Assoc. Australia and the US White House [2, 5].

classifier to block the particular page. Even a page full of images with scantily dressed people, e.g. someone's pictures from holidays on the beach, would not be classified as pornographic, since the accompanying text would not tip the classifier to that direction. On the other hand, taking images into account proved an indispensable classification aid. Although a typical pornographic Web page tends to over-advertise in text, so as to achieve a higher ranking in search-engines, there exist pornographic pages that contain very little text and many thumbnails or full-size images. These pages could not have been classified correctly without image attributes.

3 Corpus and Experimental Results

Pornographic sites are estimated to constitute around 1.5% of the Web.[5] Attempting to maintain this analogy in the corpus would result in zero learning, because the default rule of classifying everything as non-pornographic would achieve an unbeatably high performance. Instead, we assembled a corpus that consists of pornographic pages and "near-misses", the latter being non-pornographic pages that current blocking technology typically misclassifies as pornographic. Apart from being better for learning, a corpus of this kind pushes the filter to its limit when used for testing, as near-misses are much easier to confuse with pornographic pages than most ordinary pages. To collect near-misses, we queried a search engine with the first 10 keywords in the keyword blacklist of a widely used filtering tool. The engine was asked to return only pages that contained the keywords in their URLs or titles, which means that the pages would have been blocked by most current commercial filters. We then selected manually among the returned pages those that were not pornographic. This gave us 315 near-misses. Collecting pornographic pages was easier, since there exist pages with many outbound links to pornographic sites. We downloaded 534 pornographic pages, a total of 849 pages. The HTML files were pre-processed to remove common words, and the remaining words were lemmatized. Ten-fold cross-validation was then performed, ranging the number of retained attributes (attributes with highest mutual information) from 5 to 250 with a step of 5. We show only "obscene precision" and "obscene recall", as there was zero overblocking, i.e. no control page was mistakenly classified as pornographic. It appears that 100% precision can be achieved with less than 65 attributes. Recall however needs a far greater number of attributes before it can reach a comfortable 97.5%. In all cases, the

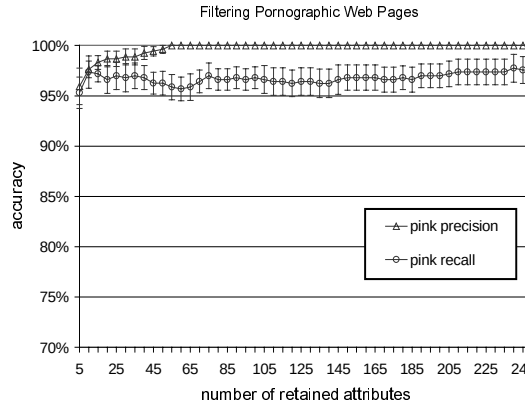

Fig. 1. Experimental results

[5] See the study cited in footnote 1.

mutual-information attribute selection retained the IMG1 attribute (section 2), but not IMG2. We examined the few (<3%) underblocked pornographic pages to find out that they were pages with all their text in clickable images, or pages with almost no text and/or images with their color balance disrupted to achieve artistic results. We are investigating techniques to trap such pages. An initial implementation of the filtering algorithm as a proxy service accessible from standard browsers has shown very promising results, with real-time processing that maintains the scores of fig. 1 and takes a few seconds, a lag that can easily go unnoticed due to Internet latency.[6]

4 Conclusions

We presented a method for detecting automatically obscene Web content, which relies on a combination of techniques from language engineering, image processing, and machine learning. Experimental results show that our method achieves nearly perfect performance on a set of hard cases in real-time.

References

1. R.O. Duda and P.E. Hart. Bayes Decision Theory. *Pattern Classification and Scene Analysis*, pp. 10–43. John Wiley, 1973.
2. D. Forsyth. Finding Naked People. Proc. of the *4th European Conference on Computer Vision*, Cambridge, England, 1996.
3. T.M. Mitchell. Bayesian Learning. *Machine Learning*, pp.154–200. McGraw-Hill, 1997.
4. J. Weinberg. Rating the Net. 19 Hastings Comm/Ent L.J. 453, 1997.
5. Clairview Internet Sheriff, An independent review. Electronic Frontiers Australia. http://www.efa.org.au/Publish/report_isheriff.html
6. Platform for Internet Content Selection (PICS). http://www.w3.org/PICS
7. Internet Content Rating Association (ICRA). http://www.icra.org
8. SafeSurf. http://www.safesurf.com/
9. P. Greenfield. *Technical Aspects of Blocking Internet Content.* National Office of the Information Economy, Australia, 1999. http://www.noie.gov.au
10. T. Minka. *An Image Database Browser that Learns from User Interaction.* MIT Media Lab TR 365, 1995.
11. T.K. Leung, M.C. Burl, and P. Perona. Finding Faces in Cluttered Scenes Using Random Labeled Graph Matching. Proc. of the *International Conference on Computer Vision*, pp. 63–644, 1995.
12. C.D. Manning and H. Schutze. Foundations of Statistical Natural Language Processing. MIT Press, 1999.

[6] See http://www.iit.demokritos.gr/filterix for relevant information.