WebGuard: Web Adult Content Detection and Filtering System

Mohamed Hammami, Ecole Centrale de Lyon, France
Liming Chen, Ecole Centrale de Lyon, France

ABSTRACT

This paper describes a Web filtering system “WebGuard,” which aims to automatically detect and filter adult content on the Web. WebGuard uses data mining techniques to classify URLs into two classes: suspect URLs and normal URLs. The suspect URLs are stored in a database, which is constantly and automatically updated in order to reflect the highly dynamic evolution of the Web. When working, WebGuard simply captures a user’s URL, matches it with the suspect URLs stored in the database and takes an appropriate action — filtering or blocking — according to the result of the analysis. We started out with a study of most existing software so as to get to know the possibilities and functionalities available on the market at the moment. This phase enabled us to better evaluate the performances of our product as it was being developed. Thus, the second phase of our work was devoted to research into the usual algorithms regarding their advantages and drawbacks. Having gathered this knowledge, we are currently implementing a system that will combine several algorithms in order to increase the software’s performance. Our preliminary results show that it can detect and filter adult content effectively.

Keywords: adult content detection; data mining; text mining; Web filtering; Web search techniques

INTRODUCTION

Nowadays, the Internet is taking an increasingly pivotal place in everyday life. Not only is the Internet community constantly growing, it is also getting younger. In fact, children each day have easier access to the Internet, which may cause socio-cultural problems. According to a study carried out in May 2000, 60% of the interviewed parents were anxious about their children navigating on the Internet, particularly because of the presence of adult material (Gralla & Kinkoph, 2001). Furthermore, according to the Forrester lookup, a company that examines operations on the Internet, online sales related to pornography adds up to 10% of the total amount of online operations (Gralla & Kinkoph, 2001). This problem concerns parents as well as companies. For example, the company Rank Xerox laid off forty
employees in October 1999 who were looking at pornographic sites during their working hours. To avoid this kind of abuse, the company installed program packages to supervise what its employees visit on the Net.

Answering this demand, some companies have proposed solutions for Web site filtering. Their products concentrate on IP-based filtering, and their classification of Web sites is mostly manual; that is to say no truly automatic classification process exists. But, as we know, the Web is a highly dynamic information source. Not only do many Web sites appear everyday while others disappear, but site content (especially links) are updated frequently. Thus, manual classification and filtering systems are largely impractical and inefficient. The ever-changing nature of the Web calls for new techniques designed to classify and filter Web sites and URLs automatically (Hammami, Tsishkou, & Chen, 2003; Hammami, Chahir, & Chen, 2003).

In this article, we propose an adult content detection and filtering system called WebGuard, which uses text analyses. Compared to other systems, WebGuard has the advantage of combining several data mining algorithms for Web site classification.

The remainder of this article is organized as follows. We started out with a study of the best-known software on the market. In the next section, the WebGuard architecture is presented. Following that, the extraction of feature vectors from Web pages is reviewed. Subsequently, the classification of URLs through data mining techniques is discussed. Finally, an experimental evaluation and comparison results are presented.

STATE-OF-THE-ART AND ANALYSIS OF THE COMPETITION

We started out with a study of best-known software on the market so as to get
Related Content

Analyzing Buffer Requirement for Wireless Enhancement of TCP in Network Mobility
[www.igi-global.com/article/analyzing-buffer-requirement-wireless-enhancement/77660?camid=4v1a](www.igi-global.com/article/analyzing-buffer-requirement-wireless-enhancement/77660?camid=4v1a)

Bridging the Digital Divide through Broadband Deployment
[www.igi-global.com/chapter/bridging-digital-divide-through-broadband/20430?camid=4v1a](www.igi-global.com/chapter/bridging-digital-divide-through-broadband/20430?camid=4v1a)
Schedulability Analysis for Real Time On-Chip Communication with Wormhole Switching
[www.igi-global.com/article/schedulability-analysis-real-time-chip/42983?camid=4v1a](www.igi-global.com/article/schedulability-analysis-real-time-chip/42983?camid=4v1a)

Subscription Policy Control Framework for IMS-Based Networks
[www.igi-global.com/chapter/subscription-policy-control-framework-ims/54010?camid=4v1a](www.igi-global.com/chapter/subscription-policy-control-framework-ims/54010?camid=4v1a)