IGI PUBLISHING **ITB14798**

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.igi-pub.com

This paper appears in the publication, **Business Data Communications and Networking: A Research Perspective** edited by **Jairo Gutiérrez**

Chapter II

A Data Mining Driven Approach for Web Classification and Filtering Based on Multimodal Content Analysis

Mohamed Hammami, Faculté des Sciences de Sfax, Tunisia

Youssef Chahir, Université de Caen, France

Liming Chen, Ecole Centrale de Lyon, France

Abstract

Along with the ever growing Web is the proliferation of objectionable content, such as sex, violence, racism, and so forth. We need efficient tools for classifying and filtering undesirable Web content. In this chapter, we investigate this problem through WebGuard, our automatic machine-learning-based pornographic Web site classification and filtering system. Facing the Internet more and more visual and multimedia as exemplified by pornographic Web sites, we focus here our attention on the use of skin color-related visual content-based analysis along with textual and structural content based analysis for improving pornographic Web site filtering. While the most commercial filtering products on the marketplace are mainly

based on textual content-based analysis such as indicative keywords detection or manually collected black list checking, the originality of our work resides on the addition of structural and visual content-based analysis to the classical textual content-based analysis along with several major-data mining techniques for learning and classifying. Experimented on a testbed of 400 Web sites including 200 adult sites and 200 nonpornographic ones, WebGuard, our Web filtering engine scored a 96.1% classification accuracy rate when only textual and structural content based analysis are used, and 97.4% classification accuracy rate when skin color-related visual content-based analysis is driven in addition. Further experiments on a black list of 12,311 adult Web sites manually collected and classified by the French Ministry of Education showed that WebGuard scored 87.82% classification accuracy rate when using only textual and structural content-based analysis, and 95.62% classification accuracy rate when the visual content-based analysis is driven in addition. The basic framework of WebGuard can apply to other categorization problems of Web sites which combine, as most of them do today, textual and visual content.

Introduction

In providing a huge collection of hyperlinked multimedia documents, Web has become a major source of information in our everyday life. With the proliferation of objectionable content on the Internet such as pornography, violence, racism, and so on, effective Web site classification and filtering solutions are essential for preventing from socio-cultural problems.

For instance, as one of the most prolific multimedia content on the Web, pornography is also considered as one of the most harmful, especially for children having each day easier access to the Internet. 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 which examines operations on the Internet, online sales related to pornography add 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 40 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.

To meet such a demand, there exists a panoply of commercial products on the marketplace proposing Web site filtering. A significant number of these products concentrate on IP-based black list 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 also frequently updated. 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).