WebGuard: A Web Filtering Engine Combining Textual, Structural, and Visual Content-Based Analysis

Mohamed Hammami, Youssef Chahir, and Liming Chen, *Member*, *IEEE Computer Society*

Abstract—Along with the ever-growing Web comes the proliferation of objectionable content, such as sex, violence, racism, etc. We need efficient tools for classifying and filtering undesirable Web content. In this paper, we investigate this problem and describe WebGuard, an automatic machine learning-based pornographic Web site classification and filtering system. Unlike most commercial filtering products, which are mainly based on textual content-based analysis such as indicative keywords detection or manually collected black list checking, WebGuard relies on several major data mining techniques associated with textual, structural content-based analysis, and skin color related visual content-based analysis as well. Experiments conducted on a testbed of 400 Web sites including 200 adult sites and 200 nonpornographic ones showed WebGuard's filtering effectiveness, reaching a 97.4 percent classification accuracy rate when textual and structural content-based analysis was combined with visual content-based analysis. 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 a 95.62 percent classification accuracy rate. 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.

Index Terms—Web classification and categorization, data mining, Web textual and structural content, visual content analysis, skin color model, pornographic Web site filtering.

1 Introduction

In providing a huge collection of hyperlinked multimedia documents, the 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, etc., effective Web site classification and filtering solutions are essential for preventing socio-cultural problems.

For instance, as some of the most prolific multimedia content on the Web, pornography is also considered as one of the most harmful, especially for children who each day have easier access to the Internet. According to a study carried out in May 2000, 60 percent of the interviewed parents were anxious about their children navigating on the Internet, particularly because of the presence of adult material [2]. Furthermore, according to the Forrester lookup, a company which examines operations on the Internet, online sales related to pornography add up to 10 percent of the total amount of online operations [2]. This problem concerns companies as well as parents. 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 this 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 is 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 [5], [6].

Automatic pornographic Web site classification is a quite representative instance of the general Web site categorization problem as it usually mixes textual hyperlinked content with visual content. A lot of research work on Web document classification and categorization has already brought to light that textual-content only-based classifiers perform poorly on hyperlinked Web documents and structural content-based features, such as hyperlinks and linked neighbor documents, help to greatly improve the classification accuracy rate [24], [32].

In this paper, we investigate the problem of pornographic Web site classification and filtering. Unlike most commercial filtering products, which are mainly based on indicative keywords detection or manually collected black list checking, our solution, WebGuard, is an automatic machine learning-based pornographic Web site classification and filtering system which relies on several major

- M. Hammami and L. Chen are with LIRIS UMR CNRS 5205, Ecole Centrale de Lyon, 36, Av Guy de Collongue, 69134 Ecully Cedex, France. E-mail: {mohamed.hammami, liming.chen}@ec-lyon.fr.
- Y. Chahir is with GREYC-URA CNRS 6072, Campus II-BP 5186, University of Caen, 14032 Caen Cedex, France. E-mail: youssef.chahir@info.unicaen.fr.

Manuscript received 23 Oct. 2004; revised 25 Apr. 2005; accepted 27 July 2005; published online 19 Dec. 2005.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-0377-1004.

data-mining algorithms associated with not only textual, structural content-based analysis, but also skin color related visual content analysis as well.

Experiments conducted on a testbed of 400 Web sites including 200 adult sites and 200 nonpornographic ones showed WebGuard's effectiveness, reaching a 97.4 percent classification accuracy rate when textual and structural content-based analysis was put together with skin color related visual content-based analysis. 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 a 95.62 percent classification accuracy rate. Based on a supervised classification with several data-mining algorithms, the basic framework of WebGuard can apply to other categorization problems of Web sites combining, as most of them do today, textual and visual content.

The remainder of this paper is organized as follows: We start out with a study of related work in Section 2. The design principle and overall architecture of WebGuard is presented in Section 3. The various features resulting from the textual and structural analysis of a Web page are described in Section 4. The skin color related visual content-based analysis is presented in Section 5. The intensive experimental evaluation and comparison results are discussed in Section 6. Some implementation issues are described in Section 7. Finally, Section 8 summarizes the WebGuard approach and presents some concluding remarks and future work directions.

2 State of the Art and Analysis of the Competition

In the literature, there exists an increasing interest in Web site classification and filtering issues. Responding to the necessity of protecting Internet access from the proliferation of harmful Web content, there is also a panoply of commercial filtering products on the marketplace. In this section, we first define some rather classical evaluation measures and describe our Web site classification testbed, the MYL test data set, which is used below to assess and compare various research work and commercial products. Then, we overview some significant research work within the field and evaluate different commercial products using MYL test data set. Finally, we conclude this state of the art section with findings from the research work overview and the analysis of commercial product competition.

2.1 MYL Test Data Set and Measures of Evaluation

A good Web content-filtering solution should deny access to adult Web sites while giving access to inoffensive ones. We thus manually collected a test data set, referred to below as the *MYL test data set*, consisting of 400 Web sites, half of them pornographic, the other half inoffensive. Manual selection of the Web sites was not straightforward because we were concerned with finding a representative selection. For instance, for the pornographic Web sites of our MYL test data set, we manually included erotic Web sites, pornographic Web sites, hack Web sites presenting pornographic nature images, and some game Web sites which, while inoffensive during the day, present illicit text and images at night.

	Assigned class	
Original class	A	B
A	$n_{A.A}$	$n_{A.B}$
B	$n_{B.A}$	$n_{B.B}$

Fig. 1. Confusion matrix for a model of two classes A and B.

The selection of nonpornographic Web sites includes the ones which may lead to confusion, in particular ones on health, sexology, fashion shows, underwear shopping sites, etc.

The performance of a classifier on a testbed can be assessed by a confusion matrix opposing the assigned class (column) of the samples by the classifier with their true original class (row). Fig. 1 illustrates a confusion matrix for a two classe model.

In this matrix, $n_{A.B}$ gives the number of samples of class A but assigned by the classifier to class B and $n_{B.A}$ the number of samples of class B but assigned to class A, while $n_{A.A}$ and $n_{B.B}$ give the number of samples correctly classified by the classifier for both classes A and B. For our pornographic Web site classification problem, A might denote the class of pornographic Web sites while B denotes that of inoffensive Web sites. Thus, a perfect Web site filtering system would produce a diagonal confusion matrix with $n_{A.B}$ and $n_{B.A}$ set to zero.

From such a confusion matrix, one can derive not only the number of times where the classifier misclasses samples but also the type of misclassification. Moreover, one can build three global indicators on the quality of a classifier from such a confusion matrix:

- **Global error rate**. $\varepsilon_{global} = (n_{A.B} + n_{B.A})/card(M)$, where *card(M)* is the number of samples in a test bed. One can easily see that the global error rate is the complement of *classification accuracy rate* or success classification rate defined by $(n_{A.A} + n_{B.B})/card(M)$.
- **A priori error rate**. This indicator measures the probability that a sample of class k is classified by the system to a class other than class k. $\varepsilon_{a\,prior}i(k) = \sum_{k \neq j} n_{k.j} / \sum_j n_{k.j}$, where j represents the different classes, i.e., A or B in our case. For instance, the a priori error rate for class A is defined by $\varepsilon_{a\,priori}(A) = n_{A.B}/(n_{A.A} + n_{A.B})$. This indicator is thus clearly the complement of the classical *recall rate*, which is defined for class A by $n_{A.A}/(n_{A.A} + n_{A.B})$.
- **A posteriori error rate**. This indicator measures the probability that a sample assigned to class k by the system effectively belongs to a class other than class k. $\varepsilon_{a\,posteriori}(k) = \sum_{k \neq j} n_{j.k} / \sum_j n_{j.k}$, where j represents the different classes, i.e., A or B in our case. For instance, the a posteriori error rate for class A is defined by $\varepsilon_{a\,posteriori}(A) = n_{B.A}/(n_{A.A} + n_{B.A})$. This indicator is thus clearly the complement of the classical *precision rate*, which is defined for class A by $n_{A.A}/(n_{A.A} + n_{B.A})$.

All these indicators are important to the assessment of the quality of a classifier. Global error rate do estimate the likelihood of error. The a priori error rate and the a posteriori error rate measure different types of errors in classification.

2.2 Related Research Work

There exist four major pornographic Web site filtering approaches, which are Platform for Internet Content Selection (PICS), URL blocking, keyword filtering, and intelligent content-based analysis [23]. PICS is a set of specification for content-rating systems which is supported both by Microsoft Internet Explorer, Netscape Navigator, and several other Web-filtering systems. As PICS is a voluntary self-labeling system rated for free by the content provider, it can only be used as the supplementary means for Web content filtering. URL blocking approach restricts or allows access by comparing the requested Web page's URL with URLs in a stored list. A *black* list contains URLs of objectionable Web sites, while a *white* list gathers permissible ones. The dynamic nature of the Web implies the necessity of constantly keeping the black list up to date, which relies in most cases on a large team of reviewers, making the human-based black list approach impracticable. The keyword filtering approach blocks access to a Web site on the basis of the occurrence of offensive words and phrases. It thus compares each word or phrase in a searched Web page with those in a keyword dictionary of prohibited words or phrases. While this approach is quite intuitive and simple, it may unfortunately easily lead to a well-known phenomena of "overblocking," which blocks access to inoffensive Web sites such as Web pages on health or sexology.

The intelligent content-based analysis for pornographic Web site classification falls in the general problem area of automatic Web site categorization and classification systems. The realization of such systems needs to rely on a machine learning process with supervised learning. For instance, Glover et al. utilized SVM in order to define a Web document classifier [32], while Lee et al. made use of neural networks to set up a Web content filtering solution [23]. The basic problem with SVM, which is revealed as very efficient in many classification applications, is the difficulty in finding a kernel function mapping the initial feature vectors into higher dimensional feature space where data from the two classes are roughly linearly separable. On the other hand, neural networks, while showing their efficiency in dealing with both linearly and nonlinearly separable problems, do not have easily understandable classification decisions.

A fundamental problem in machine learning is the design of discriminating feature vectors which relies on our a priori knowledge of the classification problem. The more simple the decision boundary is, the better is the performance of a classifier. Web documents are reputed to be notoriously difficult to classify [24]. While a text classifier can reach a classification accuracy rate between 80-87 percent on homogeneous corpora such as financial articles, it has also been shown that a text classifier is inappropriate for Web documents due to their sparse and hyperlinked structure and their diversity of Web content, which is more and more multimediatic [25]. Lee et al. proposed in their pornographic Web site classifier frequencies of indicative keywords in a Web page to judge their relevance to pornography [23]. However, they explicitly excluded URLs from their feature vector, arguing that such an exclusion should not compromise the Web page's relevance to pornography as indicative keywords contribute only a small percentage to the total occurrences of indicative keywords.

A lot of work rather emphasized the importance of Web page structure, in particular hyperlinks, to improve Web search engine ranking [26] and Web crawlers [27], discover Web communities [28], and classify Web pages [29], [30], [31], [32]. For instance, Flake et al. investigated the problem of Web community identification only based on the hyperlinked structure of the Web [25]. They highlighted that a hyperlink between two Web pages is an explicit indicator that two pages are related to one another. Starting from this hypothesis, they studied several methods and measures, such as bibliographic coupling and cocitation coupling, hub and authority, etc. Glover et al. also studied the use of Web structure for classifying and describing Web pages [32]. They concluded that the text in citing documents, when available, often has greater discriminative and descriptive power than the text in the target document itself. While emphasizing the use of inbound anchortext and surrounding words, called extended anchortext, to classify Web pages accurately, they also highlighted that the only extended anchortext-based classifier when combined with only textual content-based classifier greatly improved classification accuracy.

2.3 Analysis of Market Competition

To complete our previous overview, we also carried out a study on a set of best known commercial filtering products on the marketplace so as to get to know the performance and functionalities available at the moment. We tested the most commonly used filtering software over our MYL test data set. The six products we tested are:

1. Microsoft Internet Explorer (RSACi) [10],
2. Cybersitter 2002 [11],
3. Netnanny 4.04 [12],
4. Norton Internet Security 2003 [13],
5. Puresight Home 1.6 [14], and
6. Cyber Patrol 5.0 [15].

Most of them support PICS filtering, URL blocking, and but only keyword-based content analysis. Fig. 2 shows the results of our study. It compares the success rates of the most common software on the market today. As we can see, the success classification rate can reach 90 percent for the best of them. Interestingly enough, another independent study on the 10 most popular commercial Web-filtering systems was carried out on a data set of 200 pornographic Web pages and 300 nonpornographic Web pages and gave a similar conclusion on performance [23].

In addition to the drawbacks that we outlined in the previous section, these tests also brought to light several other issues that we discovered. A function which seems very important to users of this kind of product is the configurability of the level of selectivity of the filter. Currently, there are different types of offensive content and our study shows that, while highly pornographic sites

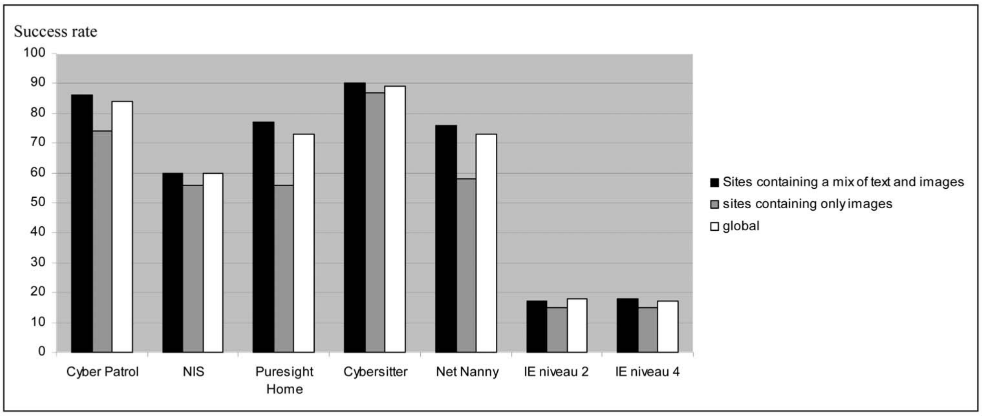

Fig. 2. Classification accuracy rates of six commercial filtering products on MYL test data set.

are well handled by most of these commercial products, erotic sites or sexual education, for instance, are unaccounted for. That is to say, they are either classified as highly offensive or as normal sites. Thus, good filters need to also be distinguished from the less good ones by their capacity to correctly identify the true nature of the pornographic or nonpornographic sites. Sites containing the word "sex" do not all have to be filtered. Adult sites must be blocked, but scientific and education sites must stay accessible.

Another major problem is the fact that all products on the market today rely solely on keyword-based textual content analysis. Thus, the efficiency of the analysis greatly depends on the word database, its language, and its diversity. For instance, we found out that a product using an American dictionary will not detect a French pornographic site.

2.4 Conclusion

To sum up, the dynamic nature and the huge amount of Web documents call for an automatic intelligent content-based approach for pornographic Web site classification and filtering. Furthermore, not only must a pornographic Web site classification and filtering solution be based on a textual content-based analysis, but it must also make use of Web page structural information such as hyperlinks, "keywords" metadata, etc. Moreover, as images are one of the key features of pornographic Web sites, a pornographic Web site filtering solution, to be fully reliable, must take into account the visual content in the classification process.

3 Principle and Architecture of WebGuard

The lack of reliability and other issues that we discovered from our previous study on the state of the art encouraged us to design and implement WebGuard with the aim to obtaining an effective Web filtering system. The overall goal of WebGuard is to make access to the Internet safer for both adults and children, blocking Web sites with pornographic content while giving access to inoffensive ones. In this section, we first sketch the basic design principles of WebGuard. Then, we introduce the fundamentals of data-mining techniques. Finally, two applications of these data-mining within the framework of WebGuard are briefly described.

3.1 WebGuard Design Principles

Given the dynamic nature of Web and its huge amount of documents, we decided to build an automatic pornographic content detection engine based on a machine learning approach which basically also enables the generalization of our solution to other Web document classification problems. Such an approach needs a learning process on an often manually labeled data set in order to yield a learned model for classification. Among various machine learning techniques, we selected a data mining approach for its comprehensibility of the learned model.

The most important step for machine learning is the selection of the appropriate features, according to our a priori knowledge of the domain, which best discriminates the different classes of the application. Informed by our previous study on the state-of-the-art solutions, we decided that the analysis of Web pages for classification should rely not only on textual content but also on its structural one. Moreover, as images are a major component of Web documents, in particular for pornographic Web sites, an efficient Web filtering solution should perform some visual content analysis.

In order to speed up navigation, we decided to use a black list whose creation and update is automatic thanks to the machine learning-based classification engine. We also decided to use a keyword dictionary as the occurrence of sexually explicit terms is an important clue for textual content and its use in the current commercial products, while reaching a classification accuracy rate up to 90 percent, showed its efficiency.

3.2 Fundamentals of Data Mining Techniques

A number of classification techniques from the statistics and machine learning communities have been proposed [7], [8], [9], [16], each having its advantages and drawbacks. But, the most important criterion for comparing classification techniques remains the classification accuracy rate. We have also considered another criterion which seems to us very important: the comprehensibility of the learned model, which leads us to a well-accepted method of classification, that is, the induction of decision trees [1], [7], [16].

A decision tree is a flowchart-like structure consisting of internal nodes, leaf nodes, and branches. Each internal node represents a decision, or test, on a data attribute and each outgoing branch corresponds to a possible outcome of the test. Each leaf node represents a class. In order to classify an unlabeled data sample, the classifier tests the attribute values of the sample against the decision tree. A path is traced from the root to a leaf node which holds the class predication for that sample.

Let Ω be the population of samples to be classified. To each sample ϖ of Ω one can associate a particular attribute, namely, its class label C. We say that C takes its value in the class of labels. For a problem of two classes c_1, c_2, one can thus write:

$$\begin{aligned} C : \Omega &\rightarrow \Gamma = \{\mathrm{c}_1, \mathrm{c}_2\} \\ \varpi &\rightarrow C(\varpi). \end{aligned}$$

For instance, c_1 might be the label representing the class of pornographic Web sites while c_2 represents the nonpornographic ones. Direct observation of $C(\varpi)$ usually is not easy; therefore, we are looking for another way φ to describe the classifier C on the basis of a combination of well-selected features. Thus, from each sample ϖ, we derive a feature vector $\mathrm{X}(\varpi) = [\mathrm{X}_1(\varpi), \mathrm{X}_2(\varpi), \ldots, \mathrm{X}_\mathrm{p}(\varpi)]$ which are also called *exogenous variables or predictive attributes*. The supervised learning consists of building a model φ from a learning data set to predict the class label of ϖ.

The process of graph construction is as follows: We begin with a learning data set and look for the particular attribute which will produce the best partition. We repeat the process for each node of the new partition. The best partitioning is obtained by maximizing the variation of uncertainty $\Im_\lambda$ between the current partition and the previous one as $I_\lambda(S_i)$ is a measure of entropy for partition S_i and $I_\lambda(S_{i+1})$ is the measure of entropy of the following partition S_{i+1}.

The variation of uncertainty is:

$$\Im_\lambda(S_{i+1}) = I_\lambda(S_i) - I_\lambda(S_{i+1}).$$

For $I_\lambda(S_i)$, we can make use of the 1) quadratic entropy or 2) Shannon entropy according to the method being selected:

$$I_\lambda(S_i) = \sum_{j=1}^{K} \frac{n_j}{n} \left(- \sum_{i=1}^{m} \frac{n_{ij}+\lambda}{n_j + m\lambda} \left(1 - \frac{n_{ij}+\lambda}{n_i + m\lambda} \right) \right), \qquad \text{(a)}$$

$$I_\lambda(S_i) = \sum_{j=1}^{K} \frac{n_j}{n} \left(- \sum_{i=1}^{m} \frac{n_{ij}+\lambda}{n_j + m\lambda} \log_2 \frac{n_{ij}+\lambda}{n_j + m\lambda} \right), \qquad \text{(b)}$$

where n_{ij} is the number of elements of class i at the node S_j with $\mathrm{I} \in \{\mathrm{c}_1, \mathrm{c}_2\}$, n_i is the total number of elements of the class i, $n_i = \sum_{j=1}^{k} n_{ij}$, n_j the number of elements of the node $S_j n_j = \sum_{i=1}^{2} n_{ij}$, n is the total number of elements, $n = \sum_{i=1}^{2} n_i$, and $m = 2$ is the number of classes $\{\mathrm{c}_1, \mathrm{c}_2\}$. λ is a variable controlling effectiveness of graph construction, it penalizes the nodes with insufficient effectives.

There exist in the literature several decision tree building algorithms, including ID3 (Induction Decision Tree) [7], C4.5, Improved C4.5, and Sipina [16]. C4.5 and Improved C4.5 [8] mainly differ from ID3 by the way of discretizing continuous values of predicative attributes, while the control and support the fusion between summits is a major specificity of Sipina [16], which stops if no changes in uncertainty occur.

3.3 Applications to Pornographic Web Site Classification and Skin Color Pixel Classification

Within the framework of WebGuard, we applied the above data mining techniques to two classification problems. The first one is, of course, pornographic Web site classification, where Ω is the population of Web sites, with c_1 representing, for instance, pornographic Web pages, while c_2 represents the nonpornographic ones.

The second problem is visual content-based analysis, namely, skin color-like pixel classification. According to such a classification, all the pixels of an image are divided into two classes: c_1 with all pixels labeled as skin color while c_2 is with all nonskin pixels.

4 Textual and Structural Content-Based Analysis

The selection of features used in a machine learning process is a key step which directly affects the performance of a classifier. Our study of the state of the art and manual collection of our test data sets grealty helped us to gain insight into pornographic Web site characteristics and to understand discriminating features between pornographic Web pages and inoffensive ones. This insight and understanding encouraged us to select both text and structure content-based features for better discrimination purposes.

4.1 Textual Content-Based Features and Keyword Dictionary

The very first and most obvious discriminating feature is the frequency of prohibited keywords within a Web page. We thus introduced *n_x_words* and *%x_words*, respectively, number of prohibited keywords and their percentage, as the first two discriminating features. However, as we evidenced in Section 2, the effectiveness and the quality of a classifier when using a keyword filtering approach depends on the nature, language, and diversity of the word database (or dictionary). We took into serious consideration the construction of this dictionary and, unlike a lot of commercial filtering products, we built a multilingual dictionary including current French, English, German, Spanish, and Italian keywords.

4.2 Structural Content-Based Features

As evidenced by the work in [32], Web structure analysis when combined with text analysis improves Web page classification and description. The structure of a Web page is introduced by tags which describe their type: hyperlinks, images, words, etc. For instance, it has been shown that hyperlinks among Web pages are important indicators of

Web communities [25]. We thus introduced *n_xxx_link*, which counts the number of "black listed" links as another feature. This attribute may describe the degree of membership of the current URL in the "black listed" community.

However, outbound links of a Web page under classification are not always classified at the time of the classification. A hyperlink has two components: the destination page and associated anchortext describing the link, which is provided by a page creator. Much as the search engine Google, which may return pages based on keywords occurring in inbound anchortext, we also use as a feature *n_x_links*, which counts the number of hyperlinks having prohibited keywords in the associated anchortext.

Similarly, it is an obvious comment that a pornographic Web site has a lot of images. Prior to a true visual content-based analysis of images, which is investigated in the next section, we also count in *n_x_images*, the number of images whose name contains sexually explicit keywords.

Other features that we introduced from analysis of various tags include:

- *n_x_meta*: The number of sexually explicit keywords in "keywords" metadata as compared to *n_meta*, the total number of words in "keywords" metadata within a Web page.
- *n_x_url*: The number of sexually explicit words in the URL of the current Web page under investigation.

4.3 Synthesis of Textual and Structural Content-Based Feature Vectors

To summarize the above, the feature vector that we proposed to characterize a Web site includes the following attributes:

1. *n_words* (total number of words on the current Web page),
2. *n_x_words* (total number of words occurring in the dictionary),
3. *n_images* (total number of images),
4. *n_x_images* (total number of images whose name has a keyword of the dictionary),
5. *n_links* (the total number of links),
6. *n_x_links* (the number of links which contain sexually explicit words),
7. *n_xxx_links* (the number of links that have been classified as sex-oriented in the black list),
8. *n_x_url* (the number of sexually explicit words in the url),
9. *n_meta* (the total number of words in "keywords" metadata),
10. *n_x_meta* (the number of sexually explicit keywords in "keywords" metadata),
11. *pcxwords* (the percentage of sexually explicit keywords),
12. *pcxmeta* (the percentage of sexually explicit keywords in "keywords" metadata),
13. *pcxlinks* (the percentage of links containing sexually explicit words), and
14. *pcximage* (the percentage of images whose name contains a sexually explicit word).

5 Skin Color Related Visual Content Analysis

It is a fact that the Web has been a major vehicle of multimedia document dissemination. A study of more than 4 million Web pages reveals that 70 percent of them contain images and there are, on average, 18.8 images per Web page [33]. Accurate Web site classification should thus take into account its visual content counterpart. The easy insight into appropriate visual content regarding a pornographic Web site is that it is obviously skin related. In this section, we describe our skin model, which also results from a supervised learning process by a data mining technique. We further improve the skin related visual content-based analysis by a region growing-based image segmentation technique. Finally, we discuss the resulting visual feature which should be used as basis for improving pornographic Web site classification.

5.1 Skin Color Modeling

Skin-color modeling is a crucial task for several applications of computer vision [4]. Problems such as face detection in video are more likely to be solved if an efficient skin-color model is constructed. Classifying skin-color pixels for Web-based adult content detection and filtering is also fundamental to the development of an accurate and reliable solution. Most potential applications of skin-color model require robustness to significant variations in races, differing lighting conditions, textures, and other factors. Given the fact that a skin surface reflects the light in a different way as compared to other surfaces, we relied, once again, on data mining techniques to define a skin color model which enables the classification of image pixels into skin ones or nonskin ones [5].

5.1.1 Skin Color Learning Data Sets

A machine learning process requires a learning data set in order to train a model. In our case, a large data set composed of tens of millions of pixels is necessary to explore various types of lighting conditions, races, etc. Two large data sets were used in our work for skin color modeling. The first one is CRL data set of skin-color and nonskin color images [21] which results from a set of 12,230 images collected by a Web crawler, consisting of 80,377,671 skin pixels and 854,744,181 nonskin pixels.

In order to further capture the lighting conditions of video images that are often encountered in pornographic Web pages, we also collected our own data set, the ECL SCIV data set, consisting of more than 1,110 skin-color images of more than 1,110 people, which resulted from 30 hours of various video sources [34]. These 1,110 skin-color images cover five races, two sexes, exterior/interior, and day/night conditions. They were manually segmented for the skin binary mask, as illustrated in Fig. 3, discriminating skin pixels from the nonskin ones.

5.1.2 Data Preparation and Data Mining-Based Learning

Let the set of pixels Ω be extracted and preprocessed automatically from training images and corresponding binary masks. We thus have a two classes classification

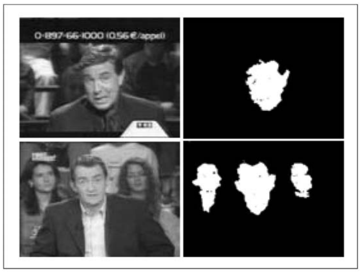

Fig. 3. Color images (left) and their corresponding skin binary mask (right).

problem, each pixel being associated with its label $C(\varpi)$: skin-color or nonskin-color.

The observation of $C(\varpi)$ is not easy because of lighting conditions, race differences, and other factors. Given that skin color is the perceived color of light reflected by a skin surface, we therefore looked for an efficient mean φ to describe class C of each pixel in different color spaces. Several color spaces have been proposed in the literature for skin detection applications. YCbCr has been widely used since the skin pixels form a compact cluster in the Cb-Cr plane. As YCbCr is also used in video coding and then no transcoding is needed, this color space has been used in skin detection applications where the video sequence is compressed [17], [18]. In [19], two components of the normalized RGB color space (rg) have been proposed to minimize luminance dependence. And, finally, CIE Lu*v* has been used in [20]. However, it is still not clear which is the color space where the skin detection performance is the best.

In our work, we computed for each pixel its representation in various normalized color spaces: RGB, HSV, YIQ, YCbCr, and CMY in order to find the most discriminative set of color axes. This leads to a feature vector V composed of 14 exogenous variables: V = [r, g, b, H, S, V, Y, I, Q, Cr, Cb, C, M, Y]. Both CRL and ECL SCIV data sets were used to generate the learning population. Associated with each pixel feature vector is its class label C: 1 for skin color and 0 for nonskin color [3]. The SIPINA [16] technique was used for training. As result, a hierarchical structure of classification rules of the type "IF...THEN..." is created. Fig. 4 illustrates some skin color pixel classification examples where the images in the middle correspond to the direct application of these decision rules.

5.2 Region Growing-Based Skin Color Image Segmentation

As we can see from the middle images in Fig. 4, there exist some pixels misclassified as skin color by our learned skin model. In order to improve the classification reliability, we further segment the skin color binary mask into skin regions by a region growing technique [35]. The basic intuition is that a skin region is a significant area with a minimum of skin pixels; otherwise, this region is a false skin-like area which needs to be filtered.

The region growing process consists of gathering neighbor pixels from a starting point on the basis of homogeneity criteria. A skin color homogeneous area within an image is a coherent area formed by all 1 pixels in its skin binary mask. More precisely, the process starts from a skin color pixel in the binary mask and tries to determine whether neighboring pixels are also skin color pixels. This process eventually leads to growing a skin-like region until no more neighboring skin color pixels can be added to the same region. In order to extract all skin-like regions, the region growing process has to be repeated for all unvisited pixels in an image.

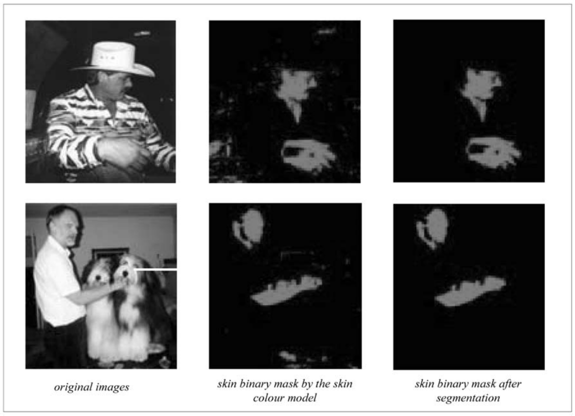
original images
skin binary mask by the skin colour model
skin binary mask after segmentation

Fig. 4. Color images (left), skin pixels classification (middle), and skin regions after segmentation (right).

All the skin-like regions are then filtered on a minimum surface criterion. Indeed, a skin-like region is considered as a skin region only when its area represents more than λ percent of the original image. The images of the last column in Fig. 4 illustrate the result of such a process. As we can see, small skin-like regions are filtered after the region growing-based segmentation process.

5.3 Visual Feature Selection

From the previous skin region segmentation technique, several strategies can be used to combine skin color related visual content feature in the Web filtering process. For instance, a straightforward one consists of extending the textual and structural content-based features V by a new skin color related feature, for instance, *%SkinPixels*, within a Web page. However, the experiments that we carried out led us rather to cascade a first Web page filter engine, below noted as WebGuard-TS, only based on textual and structural content-based features, with a second Web page engine filter engine, noted as WebGuard-V, only based on skin color related feature.

6 Experimental Results

In order to validate our Web filtering system, WebGuard was evaluated by intensive experiments. Actually, we carried out two series of tests: The first used only textual and structural content-based analysis, resulting in a first Web filter engine WebGuard-TS, while the second one took into account both textual, structural content-based analysis, and visual content-based analysis, giving birth to a cascaded version of WebGuard. This section presents the results of these experiments. However, in order to make clear the experimental conditions, we first briefly describe shortly the MYL learning data set and validation techniques and conditions.

6.1 The MYL Learning Data Sset

The creation of a database for the data mining process requires a large number of each category: In our case, 1,000 pornographic and 1,000 nonpornographic sites were added to the database. This number of 2,000 is necessary not only to simulate a good representation of Internet content, but also because we are collecting a great deal of different information. Below, we call this database of 2,000 Web sites for learning purpose as MYL learning data set.

We collected these sites manually from the Internet because we wanted our base to be as representative as possible. Within the adult sites, we find content ranging from the erotic to the pornographic and, within the nonadult sites, we find health-based information, antipornographic and anti-AIDS sites, etc.

6.2 Validation Conditions

In order to show the behavior of textual and structural content-based classification from the one integrating visual content analysis, two series of experiments were carried out: The first one used only the textual and structural content-based features, leading to the first Web filter engine WebGuard-TS, while the second one added visual content-based features in the learning and classification process, cascading WebGuard-TS and WebGuard-V.

In the first series of experiments focusing on textual and structural content-based features, we first experimented with five data mining techniques on the MYL learning data set and validated the quality of the learned model using *random error rate technique*. The stability of the learned model is further validated using *cross-validation* and *bootstrapping techniques*.

The evaluation measures are the ones defined in Section 2, namely, *global error rate*, *a priori error rate*, and *a posteriori error rate*. As global error rate is the complement of classification accuracy rate, while *a priori error rate* (respectively, *a posteriori error rate*) is the complement of the classical *recall rate* (respectively, *precision rate*). Thus, the lower the a priori error rate is achieved, the better the recall rate is. The same applies to the relationship between global error rate and global classification rate and the couple between a posteriori error rate and precision rate.

However, an efficient model on a learning data set might reveal poor performance on real data because of the so called "overfitting" phenomena or the lack of generalization ability of the learned model. A good decision tree obtained by a data mining algorithm from the learning data set should not only produce good classification performance on data already seen but also on unseen data as well. In order to ensure the performance stability of our learned model from the MYL learning data set and validated by random error rate technique, cross-validation, and bootstrapping, we thus also tested the learned model on our MYL test data set consisting of 400 Web sites. Thanks to our textual and structural content-based features, the results from these experiments showed that WebGuard-TS already outperformed the existing commercial products on the market by four or five points.

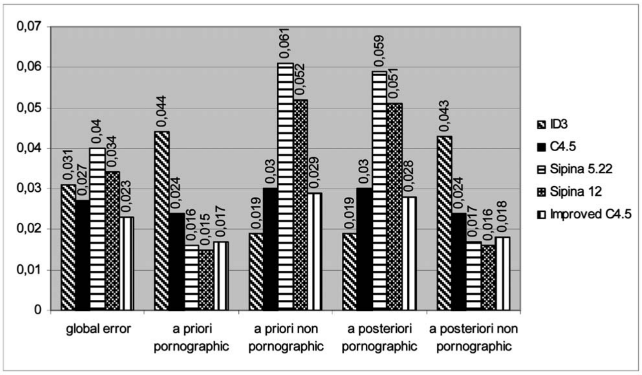

Fig. 5. Experimental results by the random error rate technique.

The second series of experiments integrated visual content-based analysis to further improve the previous experimental result. For comparison purposes with other commercial products, we carried out the experiment on the MYL test data set and WebGuard directly, cascading textual, structural, and visual analysis, reaching a classification accuracy rate up to 97.4 percent. In the following, we describe in more detail these two series of experiments.

6.3 Experiments by Textual and Structural Content-Based Classification

During this series of experiments, only textual and structural content-based features extracted from a Web page were considered. Five data mining algortithms were studied, including ID3, C4.5, Improved C4.5, Sipina with $\lambda = 5.22$ and admissibility constraint fixed at 20, noted Sipina (5.22), and Sipina with $\lambda = 12$ and admissibility of 50, noted Sipina (12). The MYL learning data set composed of 2,000 manually labeled pornographic and nonpornographic Web sites was used as the learning data set and the test data set according to the three evaluation techniques.

6.3.1 Random Error Rate Method

According to the random error rate method, the MYL learning data set was divided into two subsets: the first one for learning, composed roughly of 70 percent samples from the MYL learning data set, the other one for testing composed of the remaining samples. This process was repeated three times with random choice of the two subsets, however, keeping the ratio between the subset for learning and the one for testing. The three error rates, i.e., global error rate, a priori error rate, and a posteriori error rate, were averaged on these three experiments. The experimental results on the different data mining algorithms are depicted in Fig. 5.

As we can see in the figure, all of five data mining algorithms echoed very similar performance on the feature vector, displaying a global error rate of less than 4 percent and only 2.6 percent for the best one (Improved C4.5). There is clearly a trade-off between the a priori error rate on pornographic Web sites and nonpornographic ones and the much the same between the a posteriori error rate on pornographic Web sites and nonpornographic ones. For instance, when Sipina (12) displayed the best performance

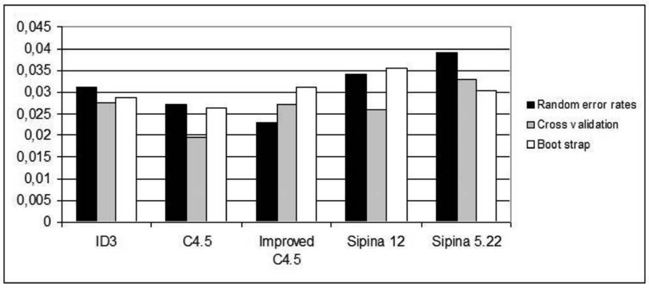

Fig. 6. Global comparison results on the three validation techniques.

on pornographic Web sites with an a priori rate of 1.5 percent, it recorded, on the other hand, an a priori error rate of 5.2 percent on nonpornographic Web sites which is the second worst performance. The same trade-off was also observed on a posteriori error rate side. When Sipina (12) achieved, with 1.7 percent, the best a posteriori error rate on nonpornographic Web sites, it recorded, on the other hand, a 5.9 percent a posteriori error rate on pornographic Web sites, which is the second worst performance among the five algorithms. It seems that the best average behavior was achieved by improved C4.5, which recorded the best global error rate of 2.3 percent, a priori error rates and a posteriori error rates ranging from 1.7 percent to 2.9 percent on pornographic Web sites and nonpornographic ones.

6.3.2 Cross-Validation and Bootstrapping

The experimental results by the random error rate technique are thus very encouraging as all five algorithms, on the basis of textual and structural content-based features, only outperform by six points the best performance displayed by commercial products that we tested. We, therefore, used the other two more systematic evaluation methods, namely, cross-validation and bootstrapping, to further confirm these results. Fig. 6 summarizes the performance of WebGuard-TS according to the three validation techniques. As we can see from the figure, even though the global error rates for the five data-learning algorithms by the random error technique are higher than the figures by the two other validation techniques, the performance obtained for each algorithm is globally in accordance as the various error rates by the three validation techniques do not differ from one another by more than 1 percent.

To conclude, the tests allow us to say that the error rates of the data mining algorithms, when using an appropriate dictionary and well-chosen parameters, are individually less than 4 percent. This rate is further lowered by a majority voting-based smoothing mechanism discussed in the following section.

6.3.3 Experimental Results on MYL Test Data Set and Classification Results Smoothing

Encouraged by the previous validation results, we then tested the five data mining algorithms on the MYL test data set which was used to compare commercial products in Section 2. The experimental results are summarized in Fig. 7. As we can see in this figure, the average global error rate by the five data mining algorithms is roughly 6 percent, which corresponds to a classification accuracy rate of 94 percent, thus four points higher than the best performance of the commercial products we evaluated. We again observe the trade-off between a priori error rate on pornographic Web sites and the nonpornographic ones and the same for a posteriori error rate between the two classes. When Sipina 12 displayed the worst performance on pornographic Web site classification, it achieved, at the same time, the best performance of a priori error rate on nonpornographic Web site classification. The best result was scored by Sipina 5.22 with only 3 percent on global error rate.

While displaying different performances, we discovered that the five data mining algorithms did not make errors in the same way. We thus decided to smooth the classification result by majority voting, leading to the first Web filter engine, WebGuard-TS. That is to say that a Web site will be classified as a pornographic one if only three of the five algorithms achieve such a classification. Using this principle

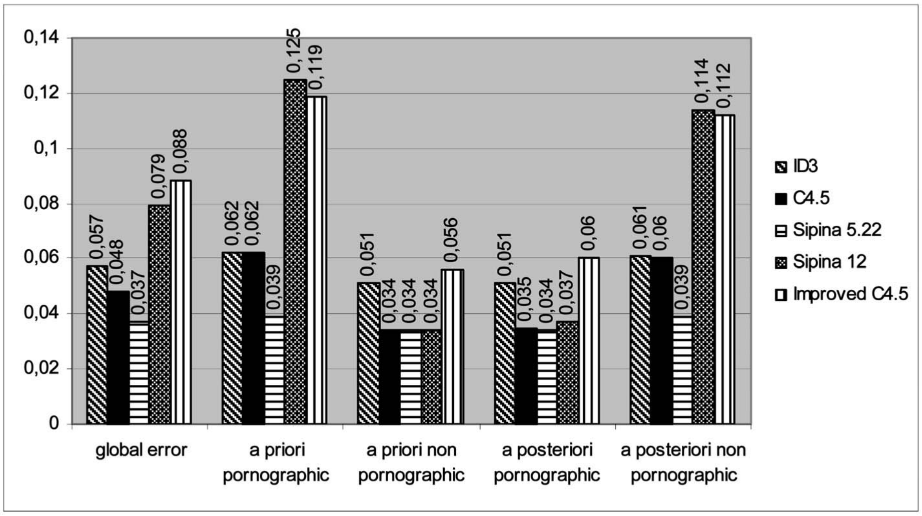

Fig. 7. Experimental results by the five algorithms on the MYL test data set.

in our experiment on the MYL test data set, WebGuard-TS further improved the performance and achieved a global error rate of 3.9 percent only.

6.4 Experiments Integrating Visual Content-Based Classification

While WebGuard-TS displays a low global error rate, its performance can be further improved by skin color related visual content-based analysis [3], [4]. From our skin region segmentation technique, several strategies can be used to combine skin color-based visual content feature. For instance, one can add a new feature, %SkinPixels, within a Web page as a new feature to the 14 features from the textual and structural content-based analysis as discussed in Section 4.3. After several experiments using our MYL learning data set, we abandoned this strategy as the visual feature was always dominated by the other textual and structural content-based features and did not appear in the final decision trees after the learning phase. However, we all know that one of the most important features of pornographic Web sites is the number of naked images or videos. These experiments led us to consider another strategy: *filters cascading*. The basic idea of filters cascading is that we first use WebGuard-TS based on textual and structural content-based analysis, which already achieved a very classification accuracy rate with a global error rate of 3.9 percent. The Web sites classified as nonpornographic are then further analyzed by our visual content-based filter that we call WebGuard-V below.

6.4.1 Preprocessing of Visual Content-Based Analysis

WebGuard-V uses the percentage of skin pixels within a Web page to discriminate pornographic ones from the nonpornographic ones. In order to obtain a discrimination threshold σ on the percentage of skin pixels, we again relied on our MYL learning data set composed of 2,000 pornographic Web sites and 2,000 nonpornographic ones. However, some precautions were needed to proceed in such a way. Indeed, we found out, on the one hand, that a lot of logo images are inserted within Web pages, distorting this discrimination threshold σ, and, on the other hand, that some "smart" pornographic content providers escape from the vigilance of keyword-based pornographic Web site filters by inserting pornographic text content into images. We thus developed a specific engine discriminating logo images from nonlogo images on the basis of image gray-level histogram analysis. For Web sites inserting pornographic content into images, another form of preprocessing is needed which consists of text detection and recognition within an image [22], [36].

Having performed all these preprocessing operations, we computed the histograms on the percentage of skin pixels from our MYL learning data set on both pornographic Web site class and the nonpornographic one. The threshold σ on the percentage of skin pixels discriminating the both classes is set to the intersection of the two Gaussian-like curves.

6.4.2 Experimental Result on WebGuard Combining Textual, Structural, and Visual Content-Based Analysis

As stated earlier, we combined textual, structural, and visual content-based analysis by cascading WebGuard-TS and WebGuard-V for Web site classification. For comparison purposes, we experimented with the cascaded WebGuard on our MYL test data set again. Fig. 8 illustrates the improvement of classification accuracy compared to the performance achieved by WebGuard-TS when only textual and structural content-based features are used.

As we can see in Fig. 8, cascaded WebGuard further improved the previous performance obtained by WebGuard-TS, achieving an a priori pornographic Web site error rate down to 2.5 percent while the global error rate was only 2.6 percent. Fig. 9 highlights the performance of WebGuard compared to other adult content detection and filtering systems, including Cyber Patrol [15], Norton Internet Security [13], Pure Sight [14], Cyber sitter [11], Net Nanny [12], and IE (Internet Explorer) [10].

This result encouraged us to further experiment cascaded WebGuard on a black list of 12,311 pornographic Web sites manually collected and classified by French Ministry of Education. While we might be a little bit disappointed by the small improvement on global error rate from 3.9 percent by WebGuard-TS to 2.6 percent by cascaded WebGuard, cascaded WebGuard significantly improved the performance, scoring a 95.55 percent classification accuracy rate by cascaded WebGuard from an 87.82 percent classification accuracy rate by WebGuard-TS.

7 Implementation Issues

7.1 Textual and Structural Content-Based Features Extraction

In order to gather a data set for the learning, testing, and feature vector of a Web site for classification, WebGuard relies on the principle of analyzing the HTML code of a Web page. We thus should be equipped with a set of functions that make it possible to read from a server and then to analyze a page. The analyzer is composed of three main functions: an http client used to connect to the Web server and retrieve the source code, an html flag analyzing function, and a content analyzer to make an initial treatment of the raw data.

7.2 Weighting System for Configurability of Objectionable Content-Level Selectivity in WebGuard

As evidenced in Section 2 from our study on several commercial filtering systems, an important functionality which most commercial products lack is the selectivity of objectionable content. Having the precise behavior of each of five data mining algorithms from our experiments on both the MYL learning and test data sets, we answered this question by setting up a weighting system combining the five data mining algorithms so that we can tune the selectivity of objectionable content by moving a threshold.

As we can see in Fig. 7, the five data mining algorithms, i.e., ID3, C4.5, SIPINA 12, Sipina 5.22, and Improved C4.5,

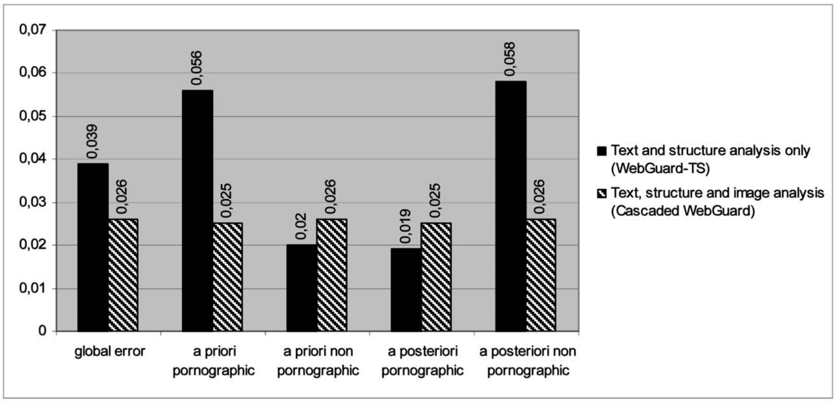

Fig. 8. Classification accuracy of cascaded WebGuard compared to WebGuard-TS.

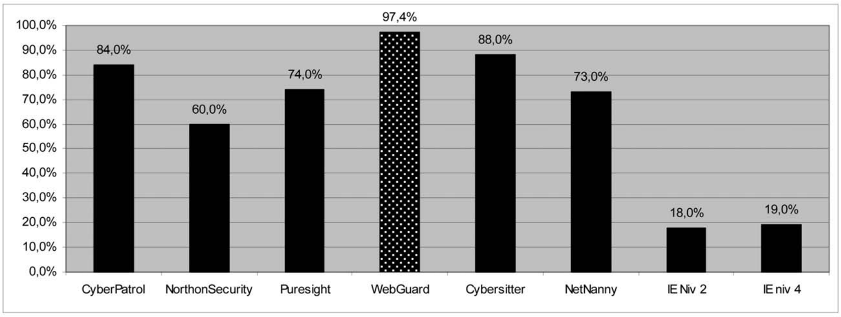

Fig. 9. Classification accuracy rate of cascaded WebGuard compared to some products.

displayed different performances on the various error rates. We might choose Sipina 5.22 for classification as it was the best algorithm achieving the minimum of various error rates on MYL test data set. Instead, we decided to combine the five data-mining algorithms together in the classification process as they produced different classification errors. However, the more reliable a data mining algorithm is, the more it should contribute in the final classification decision. We thus affected a weight γ_i associated to the classification decision χ_i of each data mining algorithm according to the formula:

$$\gamma_i = \alpha_i / (\sum_{i=1,N} \alpha_i) \text{ with } \alpha_i = (1 - (\varepsilon_i - \delta))^n,$$

where

- ε_i: the a priori error rate of the ith algorithm,
- N: the number of algorithms used for classification, here N = 5,
- n: the power in order to emphasize the difference in weight,
- σ: a threshold value that we take away from the error rate again to emphasize the difference in weight, and
- γ_i: a weight associated to the classification decision by the ith data mining algorithm.

In WebGuard, we have chosen the a priori error rates from the cross-validation results on the MYL learning data set in order to ensure that the pornographic Web sites are filtered at maximum. We might also choose other validation results as they differ in quite a few global error rates in our formula if we want to have the best-balanced behavior of our filter both on pornographic and nonpornographic classes. As the best a priori rate was a little bit more than 0.03, we set $\sigma = 0,03$. After several experiments, we fixed n = 5 as giving the best result on the MYL test data set. Using data from Fig. 7, we obtained the relative weight of the five data mining algorithms, as illustrated in Fig. 10. It shows the individual success rates of these algorithms.

As we can see in Fig. 10, the algorithm having the best performance on the MYL test data set, i.e., Sipina 5.22, received the most important weight. The final classification decision of a Web site is thus made on the basis of the following formula:

$$\gamma = i = \Sigma_{i=1,N\gamma_i\chi_I} \geq \tau \text{ with } \chi \text{ ranging from 0 to 1.}$$

τ is defined as the sensitivity to "objectionable" content. Indeed, if τ is set to 0, WebGuard is the most sensitive as a

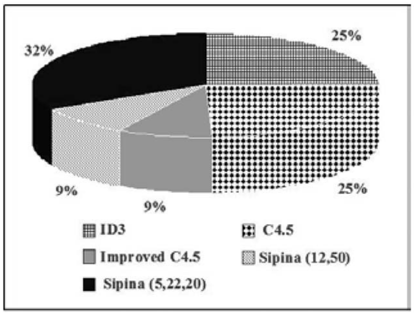

Fig. 10. Relative weights of the five data mining algorithms in the classification decision.

Web site is classified as pornographic if just one data mining algorithm does so. On the contrary, if τ is set to 1, WebGuard appears less sensitive as it classifies a Web site as pornographic only if all five data mining algorithms do so. The principle of majority voting used in our experiments for smoothing the classification results corresponds to set $\tau = 0.42$.

8 Concluding Remarks and Future Work

In this paper, we have presented WebGuard, a machine learning-based system for detecting and filtering pornographic Web pages. WebGuard combines textual and structural content-based analysis and skin color related visual content-based analysis. Our solution showed its effectiveness, scoring a 96.1 percent classification accuracy rate on our MYL test data set and a 97.4 percent classification accuracy rate when skin color relative visual-based analysis was also used in cascading WebGuard-TS and WebGuard-V. The experimental results corresponding to WebGuard-TS further confirm the discriminative power of considering structural information, in particular hyperlinks, in the classification process, as evidenced by other work [24], [25], [32]. While the slight performance improvement by cascaded WebGuard, less than two points, could suggest disregarding visual content analysis, the experimental results on another significant black list of 12,311 Web sites provided by the French Ministry of Education showed the effectiveness of skin color relative visual content-based analysis as cascaded WebGuard displayed an 95.62 percent classification accuracy rate as compared to an 87.82 percent classification accuracy rate when only textual and structural content-based analysis was used (WebGuard-TS).

We can thus summarize our major contribution by the use of both textual and structural analysis which is further combined with skin relative visual content-based analysis for Web document classification and filtering problem. However, it would be unfair to say that all the good performances were only results of textual and structural content-based analysis combined with visual content-based analysis. Actually, the dictionary of indicative keywords also played a big role in the improvement of all these performances. Currently, our dictionary contains more than 300 indicative keywords extracted from six languages. Its construction was manual, thus quite laborious and probably only possible thanks to the comprehensibility of decision trees from our data mining algorithms.

Overcoming this drawback of laborious dictionary-building is one of the directions of our future work. At this time, finding automatically discriminative indicative keywords or sentences is also typically a data mining problem. From a corpus of hyperlinked documents for one class and another one for the second class, the problem is to find indicative keywords or sentences which best discriminate the two classes.

The other direction of our future work is to leverage mutual classification capabilities both from textual and structural content-based analysis and multimedia content-based analysis. It is a fact that the Web has become more and more multimedia, including music, images, and videos. The work described in this paper suggests that Web document classification can benefit from visual content-based analysis; we also think, on the other hand, that textual and structural content-based analysis can greatly help automatic classification of images embedded within Web documents.

Acknowledgments

The authors would like to give special thanks to the anonymous reviewers for their comments which greatly contributed to the improvement of this paper. Their acknowledgment also goes to their students Florian Gérus, Nicolas Jouffroy, Audrey Licini, Fabien Perdriel, and Phan Le Bao Tuy who carried out most of the experiments described in this paper. This work was partly funded and carried out within the French national RNTL MUSE (Multimedia Search Engine) project.

References

[1] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification of Regression Trees.* Wadsworth, 1984.

[2] P. Gralla and S. Kinkoph, *Internet et les Enfants,* p. 74, CampusPress, 2001.

[3] M. Hammami, Y. Chahir, L. Chen, and D. Zighed, "Détection des Régions de Couleur de Peau dans l'Image," revue RIA-ECA, vol. 17, pp. 219-231, 2003.

[4] M. Hammami, L. Chen, D. Zighed, and Q. Song, "Définition d'un Modèle de Peau et son Utilisation pour la Classification des Images," pp. 186-197, June 2002.

[5] M. Hammami, D. Tsishkou, and L. Chen, "Data-Mining Based Skin-Color Modeling and Applications," *Proc. Third Int'l Workshop Content-Based Multimedia Indexing,* pp. 157-162, 2003.

[6] M. Hammami, Y. Chahir, and L. Chen, "WebGuard: Web Based Adult Content Detection and Filtering System" *Proc. 2003 IEEE/WIC Int'l Conf. Web Inteligence,* pp. 574-578, 2003.

[7] J.R. Quinlan, "Induction of Decision Trees," *Machine Learning,* vol. 1, pp. 81-106, 1986.

[8] J.R. Quinlan, *C4.5: Programs for Machine Learning.* Morgan Kaufmann, 1993.

[9] S.M. Weiss and C.A. Kulikowski, *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems.* Morgan Kaufman, 1991.

[10] Recreational Software Advisory Council on the Internet that became the Internet Content Rating Association (ICRA) in 1999, www.icra.org, 2005.

[11] Cybersitter 2002 Copyright © 1995-2003, Solid Oak Software, Inc., All Rights Reserved., www.cybersitter.com, 2002.

[12] Net Nanny 4.04 Copyright © 2002-2003 BioNet Systems, LLC., All Rights Reserved., www.netnanny.com, 2005.

[13] Norton Internet Security 2003 © 1995-2003 Symantec Corporation., All Rights Reserved, www.symantec.com, 2005.

[14] Puresight Home 1.6 iCognito Technologies Ltd., www.icognito.com, 2005.

[15] Cyber Patrol 5.0 © 2003 SurfControl plc., All Rights Reserved, www.cyberpatrol.com, 2005.

[16] D.A. Zighed and R. Rakotomala, "A Method for Non Arborescent Induction Graphs," technical report, Laboratory ERIC, Univ. of Lyon2, 1996.

[17] A. Albiol, L. Torres, C.A. Bouman, and E.J. Delp, "A Simple and Efficient Face Detection Algorithm for Video Database Applications," *Proc. IEEE Int'l Conf. Image Processing,* vol. 2, pp. 239-242, 2000.

[18] H. Wang and S.-F. Chang, "A Highly Efficient System for Automatic Face Region Detection in MPEG Video," *IEEE Trans. Circuits and System for Video Technology,* vol. 7, no. 4, pp. 615-628, Aug. 1997.

[19] J.G. Wang and E. Sung, “Frontal-View Face Detection and Facial Feature Extraction Using Color and Morphological Operators,” *Pattern Recognition Letters,* vol. 20, no. 10, pp. 1053-1068, Oct. 1999.

[20] M.-H. Yang and N. Ahuja, “Detecting Human Faces in Color Images,” *Proc. Int'l Conf. Image Processing,* pp. 127-130, 1998.

[21] M.J. Jones and J.M. Regh, “Statistical Color Models with Application to Skin Detection,” Cambridge Research Laboratory, CRL 98/11, 1998.

[22] S. Schüpp, Y. Chahir, A. Elmoataz, and L. Chen, “Détection et Extraction Automatique de Texte dans une Vidéo: Une Approche par Morphologie Mathématique,” pp. 73-82, MediaNet2002, 2002.

[23] P.Y. Lee, S.C. Hui, and A.C.M. Fong, “Neural Networks for Web Content Filtering,“ *IEEE Intelligent Systems,* pp. 48-57, Sept./Oct. 2002.

[24] S. Chakrabarti, B. Dom, and P. Indyk, “Enhanced Hypertext Categorization Using Hyperlinks,” *Proc. 1998 ACM SIGMOD Int'l Conf. Management of Data,* 1998.

[25] G.W. Flake, K. Tsioutsiouliklis, and L. Zhukov, “Methods for Mining Web Communities: Bibliometric, Spectral, and Flow,” *Web Dynamics,* A. Poulovassilis and M. Levene, eds. Springer Verlag, 2003.

[26] S. Brin and L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine,” *Proc. WWW7,* 1998.

[27] J. Cho, H. Garcia-Molina, and L. Page, “Efficient Crawling through URL Ordering,” *Computer Networks and ISDN Systems,* vol. 30, nos. 1-7, pp. 161-172, 1998.

[28] G.W. Flake, S. Lawrence, and C.L. Giles, “Efficient Identification of Web Communities,” *Proc. Sixth Int'l Conf. Knowledge Discovery and Data Mining (ACM SIGKDD-2000),* 2000.

[29] Y. Yang, S. Slattery, and R. Ghani, “A Study of Approaches to Hypertext Categorization,” *J. Intelligent Information Systems,* 2001.

[30] J. Fürnkranz, “Exploiting Structural Information for Text Classification on the WWW,” *Intelligent Data Analysis,* pp. 487-498, 1999.

[31] G. Attardi, A. Gulli, and F. Sebastiani, “Automatic Web Page Categorization by Link and Context Analysis,” *Proc. THAI-99, First European Symp. Telematics, Hypermedia, and Artificial Intelligence,* C. Hutchison and G. Lanzarone, eds., pp. 105-119, 1999.

[32] E.J. Glover, K. Tsioutsiouliklis, S. Lawrence, D.M. Pennock, and G.W. Flake, “Using Web Structure for Classifying and Describing Web Pages,” *Proc. WWW2002,* 2002.

[33] B. Stayrynkevitch et al., “Poesia Software Architecture Definition Document,” technical report, Poesia Consortium, Dec. 2002.

[34] E. Karpova, D. Tsishkou, and L. Chen, “The ECL Skin-Color Images from Video (SCIV) Database,” *Proc. IAPR Int'l Conf. Image and Signal Processing (ICISP '03),* pp. 47-52, 2003.

[35] Y. Chahir and L. Chen, “Efficient Content-Based Image Retrieval Based on Color Homogeneous Objects Segmentation and Their Spatial Relationship Characterization,” *J. Visual Comm. and Image Representation,* vol. 11, no. 1, pp. 302-326, Mars 2000.

[36] W. Mahdi, M. Ardebilian, L. Chen, “Text Detection and Localization within Images,” PCT/ FR03/ 02406, 31 Juillet 2002.

[37] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap.* Chapman and Hall, 1993.

Mohamed Hammami received the PhD degree in computer science from the Ecole Centrale at the Lyon Research Center for Images and Intelligent Information Systems (LIRIS) associated with the French research institution CNRS as UMR 5205. His research interest is in combining techniques of data mining and image analysis in order to resolve different classification problems such as image classification and Web site filtering. He was a staff member on the RNTL-Muse project. He has served on technical conference committees and as a reviewer for many international conferences.

Youssef Chahir received the MSc degree in fundamental computer science from the University of Technology of Compiegne in 1986, and received, in 2000 the PhD degree in signal and image processing from the Centrale Lyon. Since September 2000, he has been an assistant professor in the Departement of Computer and Information Science at the University of Caen. He is a researcher in the GREYC Laboratory at this same university. His research interests include data mining and knowledge discovery in images and video, image processing, statistical pattern recognition, and classification.

Liming Chen received the BSc degree in mathematics and computer science from the University of Nantes in 1984, the master's degree in 1986, and the PhD degree in computer science from University of Paris 6. He first served as an associate professor at the Universitè de Technologies de Compiègne, and, in 1998, he joined the Ecole Centrale de Lyon as a professor, where he has been leading an advanced research team on multimedia analysis. He is the author of more than 90 publications in the multimedia indexing field beginning in 1995. His current research interest includes cross media analysis, multimedia indexing and retrieval, face detection and recognition, and image. He is a member of the IEEE Computer Society.