Using Visual Content-based Analysis with Textual and Structural Analysis for Improving Web Filtering

MOHAMED HAMMAMI, LIMING CHEN
LIRIS UMR CNRS 5205, Ecole Centrale de Lyon 36, Av Guy de Collongue, 69134 Ecully, France
Email: {mohamed.hammami,liming.chen}@ec-lyon.fr

YOUSSEF CHAHIR
GREYC, Campus II - BP 5186 Université de Caen, 14032 Caen Cedex, France
Email: youssef.chahir@info.unicaen.fr

Received: May 19 2005; revised September 20 2005

***Abstract*— Along with the ever growing Web is 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 through WebGuard, our automatic machine learning based pornographic website classification and filtering system. Facing the Internet more and more visual and multimedia as exemplified by pornographic websites, 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 website 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 websites including 200 adult sites and 200 non pornographic 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 websites 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 websites 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 website filtering**

I. 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, etc., effective website 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 [2]. 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 [2]. 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.

To meet such a demand, there exists a panoply of commercial products on the marketplace proposing website 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 Websites and URLs automatically [5], [6].

Automatic pornographic website classification is a quite representative instance of the general website 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 only textual-content based classifier performs poorly on hyperlinked web documents and structural content-based features, such as hyperlinks and linked neighbour documents, help greatly to improve the classification accuracy rate [26], [34].

In this paper, we focus our attention on the use of skin colour related visual content-based analysis along with textual and structural content-based analysis for improving automatic pornographic website classification and filtering. Unlike the most commercial filtering products which are mainly based on 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 websites including 200 adult sites and 200 non pornographic 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 websites 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. Based on a supervised classification with several data-mining algorithms, the basic framework of WebGuard can apply to other categorization problems of websites combining, as most of them today, textual and visual content.

The remainder of this paper is organized as follows. In Section II, we first define our MYL test dataset and assessment criterion then overview related work. The design principle together with MYL learning dataset and overall architecture of WebGuard are presented in Section III. The various features resulted from textual and structural analysis of a web page and their classification performance when these features are used on MYL test dataset are described in Section IV. The skin colour modelling and skin like region segmentation are presented in section V. Based on experimental results using MYL test dataset, section VI is a comparison study of strategies for integrating skin colour related visual content-based analysis for website classification. The experimental evaluation and comparison results are discussed in Section VII. Some implementation issues including in particular image preprocessing are described in Section VIII. Finally Section VIII summarizes the WebGuard approach and presents some concluding remarks and future work directions.

II. State of the Art and Analysis of Competition

In the literature, there exists an increasing interest on website classification and filtering issue. Responding to the necessity of protecting Internet access from the proliferation of harmful Web content, there also exists a panoply of commercial filtering products on the marketplace. In this section, we first define some rather classical evaluation measures and describe our website classification testbed, MYL test dataset which is used in the subsequent 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 dataset. Finally, we conclude this state of the art section with findings from the research work overview and the analysis of commercial product competition.

TABLE I
Confusion matrix for a model of 2 classes A and B

	User sequence	
Original class	A	B
A	$n_{A.A}$	$n_{A.B}$
B	$n_{B.A}$	$n_{B.B}$

A. MYL test dataset and measures of evaluation

A good Web content-filtering solution should deny access to adult website while giving access to inoffensive ones. We thus manually collected a test dataset, named *MYL test dataset* in the subsequent, consisting of 400 websites, half of them being pornographic while the other half being inoffensive. The manual selection of these websites was a little bit tricky so as to have a good representativeness of websites. For instance, for pornographic websites of our MYL test dataset, we manually included erotic websites, pornographic websites, hack websites presenting pornographic nature images and some game websites, while inoffensive on the day, presenting illicit text and images in the night.

The selection of non pornographic websites includes the ones which may lead to confusion, in particular the ones on health, sexology, fashion parade, shopping sites on under-wear, etc.

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

In this matrix, $n_{A.A}$ 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. In our case for pornographic website classification, suppose that a web filtering engine is assessed on our MYL test dataset, we would have two classes, for instance A denoting of pornographic websites while B that of inoffensive websites. Thus, a perfect website 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*: $\epsilon_{global} = (n_{A.B} + n_{B.A})/card(M)$ 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 $(n_{A.A} + n_{B.A})/card(M)$;

- *A priori error rate*: this indicator measures the probability that a sample of class k is classified by the system to other class than class k. $\epsilon_{apriori}(k) = \sum_{j \neq k} n_{k.j} / \sum_j n_{k.j}$ 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 $\epsilon_{apriori}(A) = n_{A.B}/(n_{A.A} + n_{A.B})$. 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 class k. $\epsilon_{apriori}(k) = \sum_{j \neq k} n_{j.k} / \sum_j n_{j.k}$ 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 $\epsilon_{aposteriori}(A) = n_{B.A}/(n_{A.A} + n_{B.A})$. 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 on the assessment of the quality of a classifier. When global error rate gives the global behaviour of the system, a priori error rate and a posteriori error rate tell us more precisely where the classifier is likely to commit wrong results.

B. Related research work

There exist four major pornographic website filtering approaches which are Platform for Internet Content Selection (PICS), URL blocking, keyword filtering and intelligent content-based analysis [25]. 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-labelling system freely rated by content provider, it can only be used as supplementary mean for web content filtering. URL blocking approach restricts or allow access by comparing the requested Web pages URL with URLs in a stored list. A *black* list contains URLs of objectionable Websites while a *white* list gathers permissible ones. The dynamic nature of Web implies the necessity of constantly keeping to date the black list which relies in the most cases on large team of reviewers, making the human based black list approach impracticable. Keyword filtering approach blocks access to website 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 of 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 websites for instance web pages on health or sexology.

The intelligent content-based analysis for pornographic website classification falls in the general problem of automatic website categorization and classification systems. The elaboration of such systems needs to rely on a machine learning process with a supervised learning. For instance, Glover *et al*. utilized SVM in order to define a web document classifier [34] while Lee *et al*. made use of neural networks to set up a web content filtering solution [25]. The basic problem with SVM which reveals to be very efficient in many classification applications is the difficulty of 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 its efficiency in dealing with both linearly and non linearly separable problems, are not easy to understand its classification decision.

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 [26]. While a text classifier can reach a classification accuracy rate between 80–87% on homogeneous corpora such as financial articles, it has also been shown that a text classifier is inappropriate for web documents due to sparse and hyperlinked structure and its diversity of web contents more and more multimedia [27]. Lee et al. proposed in their pornographic website classifier frequencies of indicative keywords in a web page to judge its relevance to pornography [25]. However, they explicitly excluded URLs from their feature vector, arguing that such an exclusion should not compromise the webpage's relevance to pornography as indicative keywords contribute only a small percentage to the total occurrences of indicative keywords.

A lot of work emphasized rather the importance of web page structure, in particular hyperlinks, to improve web search engine ranking[28][40] and web crawlers [29], discover web communities [30], and classify web pages[31]–[34]. For instance, Flake *et al*. investigated the problem of Web community identification only based on the hyperlinked structure of the Web [27]. They highlighted that a hyperlink between two Web pages is an explicit indicator that two pages are related to one another. Started from this hypothesis, they studied several methods and measures, such as bibliographic coupling and co-citation coupling, hub and authority, etc. Glover *et al*. also studied the use of Web structure for classifying and describing Web pages [34]. 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 the classification accuracy. However, none of these work propose to take into account the visual content for web classification.

C. 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 dataset.

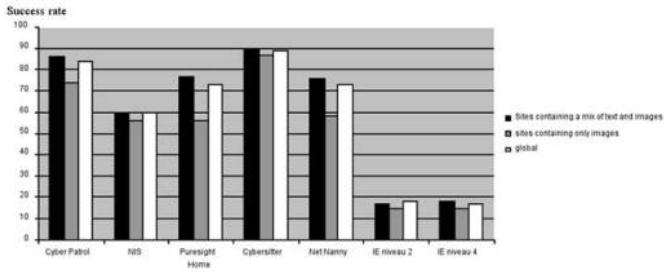

Fig. 1. Classification accuracy rates of 6 commercial filtering products on MYL test dataset.

The six products we tested are (1) Microsoft Internet Explorer (RSACi) [12], (2) Cybersitter 2002 [13], (3) Netnanny 4.04 [14], (4) Norton Internet Security 2003 [15], (5) Puresight Home 1.6 [16] and (6) Cyber Patrol 5.0 [17].

Most of them support PICS filtering, URL blocking and but only keyword-based content analysis. Figure 1 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% for the best of them. Interestingly enough, another independent study on the most 10 popular commercial Web-filtering systems was driven on a dataset of 200 pornographic web pages and 300 non pornographic web pages and gave similar conclusion on performance [25].

In addition to 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. Actually there are different types of offensive content and our study shows that, while highly pornographic sites are well handled by the 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 are to be distinguished from the less good ones also by their capacity to correctly identify the true nature of the pornographic or non-pornographic 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.

D. Overview outlines

To sum up, the most commercial filtering products are mainly based on indicative keywords detection or manually collected black list checking while the dynamic nature and the huge amount of Web documents call for an automatic intelligent content-based approach for pornographic website classification and filtering. Furthermore, if many related research work suggest with reason the importance of structural information, such as hyperlinks, "keywords" metadata, etc., for website classification and categorization, they do not take into account the visual content while the Internet has become more and more visual as exemplified by the proliferation of pornographic websites. A fully efficient and reliable pornographic website classification and filtering solution thus must be automatic system relying on textual and structural content-based analysis along with visual content-based analysis.

III. 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 Internet safer for both adults and children, blocking websites with pornographic content while giving access on inoffensive ones. In this section, we first sketch the basic design principles of WebGuard; then, we introduce the fundamentals of data-mining techniques which are used as the basic machine learning mechanism in our work. Following that, two applications of these data-mining within the framework of WebGuard are shortly described; Finally, the MYL learning dataset are presented.

A. 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 problem. Such an approach needs a learning process on a often manually labelled dataset in order to yield a learnt model for classification. Among various machine learning techniques, we selected data-mining approach for its comprehensibility of the learnt model.

The most important step for machine learning is the selection of the appropriate features, according to the a priori knowledge of the domain, which best discriminate 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 page 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 websites, an efficient web filtering solution should perform some visual content analysis.

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 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%, showed its efficiency.

B. Fundamentals of data-mining techniques

A number of classification techniques from the statistics and machine learning communities have been proposed [9]–[11], [18]. As highlighted in the studies [41], [42] on the use of data mining techniques in various applications, each one has

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], [9], [18], [43].

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 $\overline{\omega}$ 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:

$$C : \Omega \rightarrow \Gamma = \{c_1, c_2\}$$
$$\overline{\omega} \rightarrow C(\overline{\omega})$$

For instance, c_1 might be the label representing the class of pornographic websites while c_2 the non pornographic ones. Direct observation of $C(\overline{\omega})$ 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 $\overline{\omega}$) we derive a feature vector $\mathrm{X}(\overline{\omega}) = [\mathrm{X}_1(\overline{\omega}), \mathrm{X}_2(\overline{\omega}), ..., \mathrm{X}_p(\overline{\omega})]$ which are also called *exogenous variables* or *predictive attributes*. The supervised learning consists of building a model φ from a learning dataset to predict the class label of $\overline{\omega}$.

The process of graph construction is as follows: We begin with a learning dataset 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 the quadratic entropy (a) or Shannon entropy (b) according to the method being selected:

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

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

where n_{ij} is the number of elements of class i at the node S_j with $\mathrm{I} \in \{c_1, c_2\}$; n_i is the total number of elements of the class i, $n_i = \sum_{j=1}^{k} n_{nj}$; n_j is the number of elements of the node S_j, $n_j = \sum_{i=1}^{2} n_{ij}$; n is the total number of elements, $n = \sum_{i=1}^{2} n_i$; $m = 2$ is the number of classes c_1, c_2. λ is a variable controlling effectiveness of graph construction, it penalizes the nodes with insufficient effectives.

There exists in the literature several decision tree building algorithms, including ID3 (Induction Decision Tree) [9], C4.5, Improved C4.5 and Sipina [18]. C4.5 and Improved C4.5 [10] mainly differ from ID3 by the way of discretising continuous values of predicative attributes while the control and support the fusion between summits is a major specificity of Sipina [18] which stops if no changes in uncertainty occur.

C. Applications to pornographic website classification and skin colour pixel classification

Within the framework of WebGuard, we applied the above data-mining techniques to two classification problems [8]. The first one is of course pornographic website classification where Ω is the population of websites with c_1 representing for instance pornographic web pages while c_2 the non pornographic ones. To each website s is thus associated a class attribute $C(s)$ which takes two values, for instance 0 for adult websites while 1 for normal ones:

$$C : \Omega \rightarrow \Gamma = \{Adult, Non\_Adult\}$$
$$s \rightarrow C(s)$$

As direct observation of C is not easy, we look for setting up a prediction model ψ which describes the class attribute $C(s)$ on the basis of a well selected feature vector: $X(s) = [X_1(s), X_2(s), ..., X_p(s)]$ that we can extract from an automatic analysis of a website content s. The supervised learning by the aforementioned data-mining algorithms consists of building a decision tree based model ψ from a learning dataset to predict the class attribute of each website s. As illustrated by Fig. 2, a whole data mining process consists of three major steps:

- Selection and pre-processing step which consists of select the features which best discriminate classes and extract the feature vectors from the learning dataset;
- Data-mining step which looks for a synthetic and generalizable model by the use of various algorithms;
- Evaluation and validation step which consists of assessing the quality of the learnt model on the learning dataset but preferably other dataset.

In much a similar way as we will detail in the following sections, the second problem is visual content-based analysis, namely skin colour like pixel classification. According to such a classification, all the pixels of an image are divided into two classes: c_1 with all pixels labelled as skin colour while c_2 with all non skin pixels.

D. The MYL learning dataset

From the previous sections, we see that the data-mining process for pornographic website classification requires a representative learning dataset consisting of a significant set

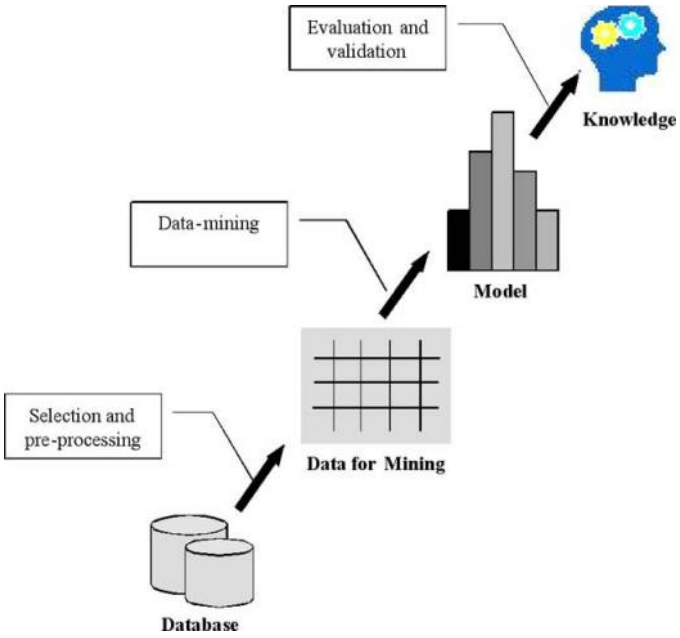

Fig. 2. Data-mining process from data.

of manually classified websites. In our case, we collected a large number of websites in each category: 1000 pornographic websites and 1000 non-pornographic ones. This number of 2000 is necessary not only to simulate a good representation of Internet content but also because we are collecting a great deal of different information. In the subsequent, we will call this database of 2000 websites for learning purpose as MYL learning dataset.

We have 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 non-adult sites we find health-based information, anti-pornographic and anti-AIDS sites, etc.

IV. 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 datasets helped us a lot to gain intuition on pornographic website characteristics and to understand discriminating features between pornographic web pages and inoffensive ones. These intuition and understanding suggested us to select both textual and structural content-based features for better discrimination purpose. The prediction model learnt from MYL learning dataset using these textual and structural content-based features leads to WebGurad-TS, our first version of pornographic website classification and filtering solution displaying high filtering effectiveness as evidenced by the experiments on MYL test dataset.

A. Textual content-based features and keyword dictionary

The very first and evident discriminating feature is 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 keyword filtering approach depend on the nature, language, and diversity of the word database (or dictionary). We did take care of this dictionary elaboration, and unlike a lot of commercial filtering products we built a multilingual dictionary including currently French, English, German, Spanish and Italian keywords.

B. Structural content-based features

As evidenced by the work in [34], 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 indicator of web communities [27]. We thus introduced n_xxx_link that 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 feature n_x_links which counts the number of hyperlinks having prohibited keywords in the associated achortext.

Similarly, it is an evident intuition that a pornographic website 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.

C. Experimental results on MYL test dataset and classification results smoothing

To summarize the above, the feature vector that we proposed to characterize a website includes the following attributes: n_words (total number of words on the current Web page); n_x_words (total number of words occurring in the dictionary), n_images (total number of images), n_x_images (total number of images whose name has a keyword of the dictionary), n_links (the total number of links), n_x_links (the number of links which contain sexually explicit words), n_xxx_links (the number of links that have been classified as sex-oriented in the black list), n_x_url (the number of sexually explicit words in the url), n_meta (the total number of words

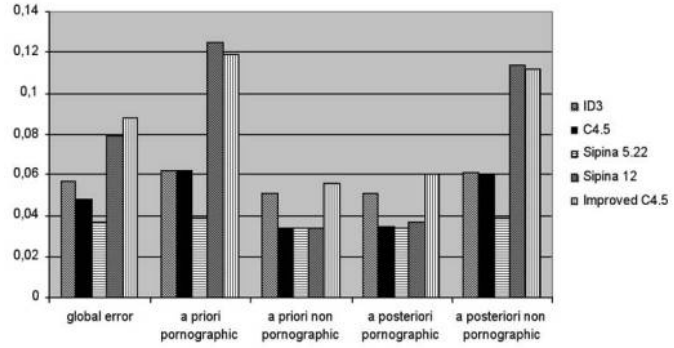

Fig. 3. Experimental results by the five algorithms on MYL test dataset only using textual and structural features.

in "keywords metadata), n_x_meta (the number of sexually explicit keywords in "keywords" metadata), $pcxwords$ (the percentage of sexually explicit keywords), $pcxmeta$ (the percentage of sexually explicit keywords in "keywords" metadata), $pcxlinks$ (the percentage of links containing sexually explicit words), $pcximage$ (the percentage of images whose name contains a sexually explicit word).

We experimented the five data-mining algorithms on MYL test dataset which was used to compare commercial products in section 2. The experimental results are summarized by Fig. 3. As we can see from this figure, the average global error rate by the five data-mining algorithms are roughly 6% which corresponds to a classification accuracy rate of 94%, thus 4 points higher that the best performance of the commercial products that we evaluated. We observe the tradeoff between a priori error rate on pornographic websites and the non pornographic ones and the same for posteriori error rate between the two classes. When Sipina 12 displayed the worst performance on pornographic website classification, it achieved on the same time the best performance of a priori error rate on non pornographic website classification. The best result was scored by Sipina 5.22 with only 3% on global error rate.

While displaying different performance, 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 a majority voting, leading to the first web filter engine, WebGuard-TS. That is to say that a website will be classified as pornographic one only if three of the five algorithms achieve such a classification. Using this principle in our experiment on MYL test dataset, WebGuard-TS further improved the performance and achieved a global error rate of 3.9% only.

V. Skin Region Segmentation

While WebGuard-TS only using textual and structural content-based analysis displayed high effectiveness for pornographic website classification and filtering, we will show that the addition of visual content-based analysis can further improve the performance. It is a fact that Web has been a major vehicle of multimedia document dissemination. A study on more than 4 million web pages reveals that 70% of them contain images and there are on average 18.8 images per web page [35]. Accurate website classification should thus take into account its visual content counterpart. While content-based image retrieval (CBIR) has been the focus of a lot of works in the literature [37], [44], the easy intuition of appropriate visual content regarding pornographic website is evidently skin related [7], [45]. 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. The study of strategies in order to take into account the resulting skin related visual feature for improving pornographic website classification will be discussed in the next section.

A. Skin colour modelling

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 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 colour model which enables the classification of image pixels into skin ones or non skin ones [5].

1) *Skin colour learning datasets*: A machine learning process requires a learning dataset in order to train a model. In our case, large datasets composed of tens of millions of pixels are necessary to explore various types of lighting conditions, races, etc. Two large datasets were used in our work for skin colour modelling. The first one is CRL dataset of skin-color and non-skin color images [23] 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 non skin 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 dataset, ECL SCIV dataset, consisting of over 1110 skin-color images of more than 1110 people, which resulted from 30 hours of various video sources [36]. These 1110 skin-color images cover 5 races, 2 sexes, exterior/interior and day/night conditions. They were manually segmented for skin binary mask, as illustrated in Fig. 4, discriminating skin pixels from the non skin ones.
2) *Data preparation and data-mining based learning*: Let the set of pixels Ω be extracted and pre-processed automatically from training images and corresponding binary masks. We thus have a two classes classification problem, each pixel $\overline{\omega}$ being associated with its label $C(\overline{\omega})$: skin-color or non skin-color.
   The observation of $C(\overline{\omega})$ is not easy, because of lighting conditions, race differences and other factors. Given that skin colour is perceived colour of light reflected by a

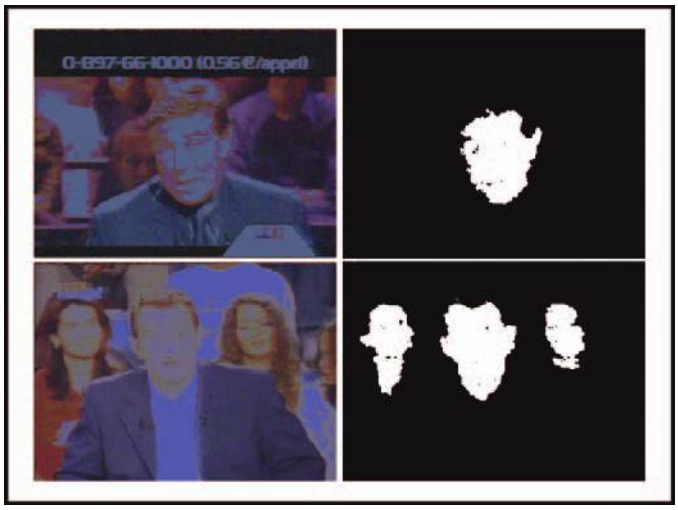

Fig. 4. Color images (left) and their corresponding skin binary mask (right).

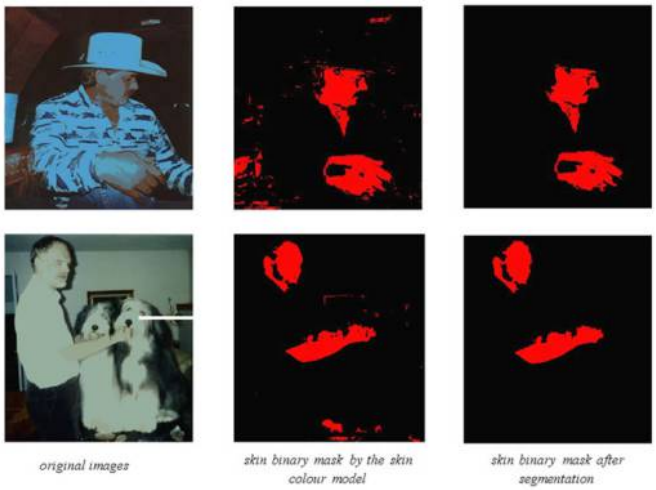

Fig. 5. Colour images (left), skin pixels classification (middle) and skin regions after segmentation (right).

skin surface, we therefore looked for an efficient mean φ to describe class C of each pixel in different colour 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 [19], [20]. In [21] 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 [22]. 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, 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 datasets were used to generate the learning population. Associated with each pixel feature vector is its class label C: 1 for skin colour and 0 for non skin colour [3]. SIPINA [18] technique was used for training. As result, a hierarchical structure of classification rules of the type "IF...THEN..." is created. Figure 5 illustrates some skin colour pixel classification examples where the images in the middle correspond to the direct application of these decision rules.

B. Skin colour region growing

As we can see from the middle images in Fig. 5, there exist some pixels misclassified as skin colour by our learnt skin model. To improve the classification reliability, we further segment the skin colour binary mask into skin regions by a region growing technique [37]. The basic intuition is that a skin region is a significant area with a minimum of skin pixels, otherwise this region is a noisy 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 colour pixel in the binary mask, and tries to determine whether neighboring pixels are also skin colour pixels. This process eventually leads to grow a skin like region until no more neighboring skin colour 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.

All the skin like regions are then filtered on a minimum surface criterion. Indeed, a skin like region is considered as skin region only when its area represents more than $\lambda\%$ of the original image. The images of the last column in Fig. 5 illustrate the result of such a process. As we can see, small skin like regions are filtered after the region growing based segmentation process.

VI. Stragety Study for Integrating Skin Colour Related Visual Content-based Analysis

From the previous skin region segmentation technique, two strategies can be used to combine skin colour related visual content feature in the web filtering process. A first straightforward strategy, that we can call *strategy of homogeneity* in the subsequent, consists of extending the fourteen textual and structural content-based features by a new skin colour related feature, for instance *%SkinPixels* within a webpage. We thus need to make a new training on MYL learning dataset using the new feature vector in order to obtain a new prediction model for classifying and filtering websites. The second strategy consists of cascading WebGuard-TS, our website filtering engine only based on textual and structural content-based features, with a second website filtering engine, noted as WebGuard-V, only based on skin colour related feature. Experiments were carried out on MYL test dataset in order to evaluate and compare these two strategies.

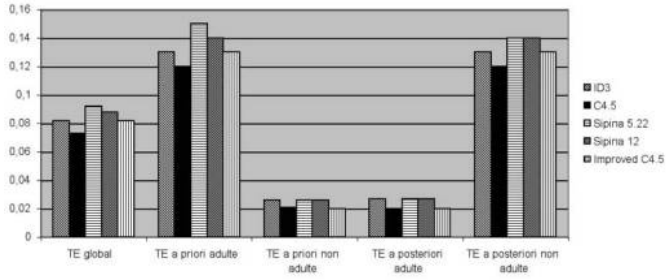

Fig. 6. Experimental results by the strategy of homogeneity on MYL test dataset.

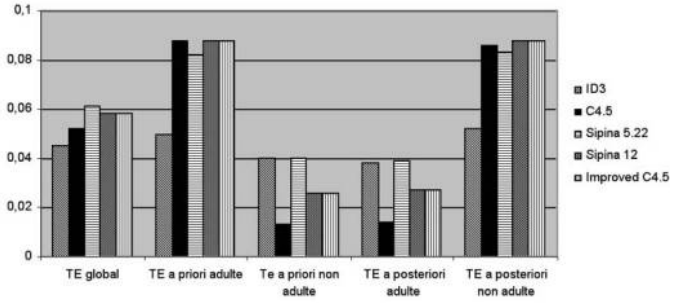

Fig. 7. Experimental results of the cascading strategy - WebGuard-V (% pornographic images).

A. Strategy of homogeneity

According to this strategy, the visual content of a webpage is used along with the textual and structural content for better discriminating pornographic websites from the normal ones. We thus proposed to extend the fourteen textual and structural features as described in the previous section by the following eleven other features related to visual content of a webpage:

- Number of adult images within a webpage;
- Percentage of adult images within a webpage;
- Number of adult images whose name contains a keyword of the dictionary;
- Percentage of adult images whose name contains a keyword of the dictionary;
- Number of logos within a webpage;
- Percentage of logos within a webpage;
- Number of logos whose name contains a keyword of the dictionary;
- Percentage of non skin pixels within a webpage;
- Number of normal images;
- Percentage of normal images ;
- Percentage of skin pixels within a webpage.

Figure 6 summarizes the experimental results which are obtained by the five data-mining algorithms on MYL test dataset using this new feature vector.

As we can see from the figure, the strategy of homogeneity for visual content integration displayed bad performance as compared to WebGuard-TS which only used textual and structural features. This is explained by the fact that visual features such as skin pixels within images are not features having the same granularity as compared to textual and structural features.

B. Strategy of cascading

The basic idea is the following. As WebGuard-TS displayed very good performance, we can consider a strategy of cascading which applies first WebGuard-TS and then an engine only based on skin colour related visual content analysis, called in the subsequent WebGuard-V, which further examines normal websites classified by WebGuard-TS. However, there are again two variants that we study in the subsequent.

1) *The first variant of cascading strategy using percentage of pornographic images*
The first variant of our cascading strategy, that we note WebGuard-V (% Pornographic Images), integrates skin related visual analysis by computing a percentage of potential pornographic images within a webpage. When this percentage exceeds a threshold, then the website under analysis is classified as pornographic. According to such as variant, we need to first determine a threshold setting the percentage of skin colour pixels within an image from which the image can be considered as potentially pornographic. For this purpose, we collected a dataset of 6000 images extracted from the Internet. The analysis of this dataset gives an average of 18% of skin pixels per image for 4000 non pornographic images which include 700 portray images, and an average of 45% of skin pixels per image for the remaining 2000 pornographic images of our dataset. After several experiments, we chose 26% as the threshold of skin colour pixel percentage from which an image is classified as potentially pornographic. According to such a threshold, 78% of pornographic images are effectively classified as pornographic while misclassification occurred for 23% of non pornographic images.
Figure 7 displays the various performances when this strategy, coupled with WebGuard-TS, is experimented on MYL test dataset. As we can see from the figure, the results are much better as compared to the ones obtained when using the first strategy of homogeneity. They even improve slightly the performance achieved by WebGuard-TS.
2) *The second variant of cascading strategy using percentage of total skin colour pixels within a webpage*
The second variant of the cascading strategy, that we note as WebGuard-V (% skin colour), proposes to consider rather the percentage of the total skin pixels from all the images within a website. When this percentage exceeds a threshold, the webpage is then considered as pornographic. Using all the images included within our MYL learning dataset, we obtained the two curves of skin colour pixel percentage depicted in Fig. 8 for normal websites and pornographic ones. As we can see, these two curves depict a gaussian behaviour and the best threshold for discriminating pornographic websites from the normal ones is 24%. Using such as threshold, the performance that we achieved was very similar to the one by the previous variant using percentage of potential pornographic images.

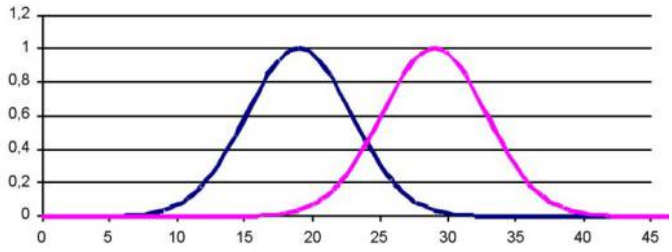

Fig. 8. Gaussian curves of skin colour pixel percentage (left for normal websites and right for pornographic ones).

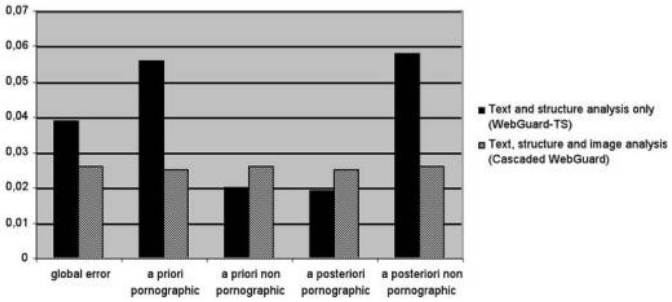

Fig. 9. Classification accuracy of cascaded WebGuard compared to WebGuard-TS.

When carrying out a detailed analysis on these results, we discovered that many images are actually logo images or images containing only text. Once eliminated these logo or text images by an automatic process that we will explain in the section on implementation issue, we obtained a threshold of 34% of skin pixel percentage which discriminates much better pornographic websites from normal ones as we will see in the next section

C. Comparison synthesis

The previous comparison study on visual content integration strategies led us to choose the cascading strategy with the second variant using the total percentage of skin pixels within a webpage. In summary, if we designate WebGuard-V our website classification engine only based on skin colour related analysis, our global website classification engine, that we note as WebGuard-TSV, consists of applying first WebGuard-TS for classification then WebGuard-V for further checking, thus cascading WebGuard-V over WebGuard-TS which only relies on textual and structural content-based analysis.

VII. Experimental Results

In order to highlight the performance improvement by considering skin colour related content analysis [3], [4], we carried our a first experiment which compares, on the basis of MYL test dataset, WebGuard-TSV with WebGuard-TS which already displayed a low global error rate. Figure 9 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 from Fig. 9, cascaded WebGuard-TSV further improved the performance obtained by WebGuard-TS, achieving a priori pornographic website error rate down to 2.5%

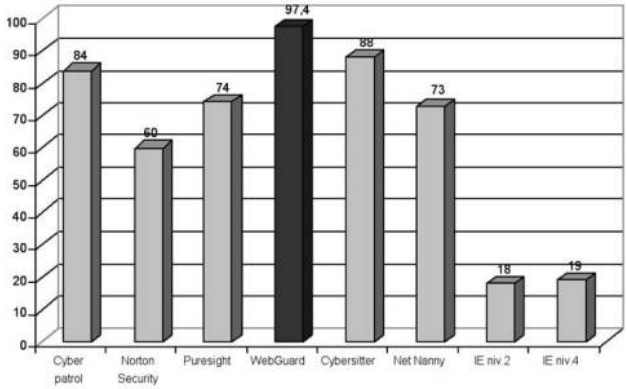

Fig. 10. Classification accuracy rate of cascaded WebGuard-TSV compared to some products.

TABLE II
CLASSIFICATION RESULTS BY WEBGUARD-TS ON THE BLACK LIST FROM THE FRENCH MINISTRY OF EDUCATION (CLASSIFICATION ACCURACY RATE: 87.82%)

WebGuard-TS	Classification results	
Website category	Adult websites	Normal websites
Adult websites	5819	1046
Normal websites		1723
Vanished websites	3723	

while the global error rate was only 2.6%. Figure 10 further highlights the performance of WebGuard-TSV compared to other adult content detection and filtering systems, including Cyber Patrol [17], Norton Internet Security [15], Pure Sight [16], Cyber sitter [13], Net Nanny[14], IE (Internet Explorer) [12].

This result encouraged us to further experiment cascaded WebGuard on a black list of 12311 pornographic websites manually collected and classified by French Ministry of Education. Tables 2 and 3 display classification results. While we might be a little bit disappointed by the slight improvement on global error rate from 3.9% by WebGuard-TS to 2.6% by cascaded WebGuard-TSV, WebGuard-TSV improved significantly the performance, as we can see from these tables, scoring a 95.55% classification accuracy rate by cascaded WebGuard from a 87.82% classification accuracy rate by WebGuard-TS.

When carrying out this experiment, we first discovered that 3723 websites have disappeared, illustrating the extreme dynamic nature of the Internet. Furthermore, according to these tables, 1723 websites were classified as normal ones both by WebGuard-TS and WebGuard-TSV. After a laborious

TABLE III
CLASSIFICATION RESULTS BY WEBGUARD-TSV ON THE BLACK LIST FROM THE FRENCH MINISTRY OF EDUCATION (CLASSIFICATION ACCURACY RATE: 95.62%)

WebGuard-TS	Classification results	
Website category	Adult websites	Normal websites
Adult websites	6489	376
Normal websites		1723
Vanished websites	3723	

phase of manual checking, we effectively discovered that these websites are normal ones. Their content had simply been changed when our experiment was carried out. These discovers reinforce again the necessity for automatic website classification engines.

VIII. Implementation Issues

WebGuard website classification engine was implemented in C++ on a PC platform while the rule based knowledge by all the data-mining process described in this paper was generated using a Sipina platform. Some other miscellaneous implementation issues include extraction of texual and structural content-based features, image preprocessing and WebGuard configurability. They are described in the subsequent.

A. Textual and structural content-based features extraction

In order to gather dataset for learning, testing and feature vector of a website 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 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.

B. Logo image discrimination

WebGuard-V uses the percentage of skin pixels within a web page to discriminate pornographic ones from the non pornographic ones. In order to obtain a discrimination threshold δ on percentage of skin pixels, we relied again on our MYL learning dataset composed of 2000 pornographic websites and 2000 non pornographic ones. However, some precautions were needed in proceeding in such a way. Indeed, we found out, on the one hand, that a lot of logo images are inserted within web pages as illustrated by Fig. 11, distorting this discrimination threshold δ, and on the other hand that some "smart" pornographic content providers escape from the vigilance of keyword-based pornographic website filters by inserting pornographic text content into images. We thus developed a specific engine discriminating logo images from non logo ones on the basis of image grey level histogram analysis. As logo images tend to have very few picks as compared to non logo images, we built a very simple discrimination decision as follows: once computed the grey level histogram of an image under analysis, we count the number of picks; if the number of picks exceeds an empirical threshold, the image under analysis is considered as non logo image.

Once performed all this pre-processing, we computed the histograms on percentage of skin pixels from our MYL learning dataset both on pornographic website class and the non pornographic one. The threshold δ on percentage of skin pixels discriminating the both classes is set to the intersection of the two Gaussian like curves and the optimal value was found to be 34%.

For websites inserting textual pornographic content into images, another pre-processing is needed which consists of text detection and recognition within an image [24], [38].

C. 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 lacks to most of commercial products is the selectivity of objectionable content. Having precise behaviour of each of five data-mining algorithms from our experiments on both MYL learning and test dataset, 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 from Fig. 3, the five data-mining algorithms, i.e. ID3, C4.5, SIPINA 12, Sipina 5.22 and Improved C4.5, displayed different performances on the various error rate. We might choose Sipina 5.22 for classification as it was the best algorithm achieving minimum of various error rates on MYL test dataset. 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:

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

where

- γ_i: the a priori error rate of the i-th algorithm;
- N: the number of algorithms used for classification, here N=5;
- n: the power in order to emphasize the difference in weight;
- δ: a threshold value that we take away from the error rate again to emphasize the difference in weight;
- γ_i: the classification decision by the i-th data-mining algorithm on a website, 0 for non pornographic class while 1 for pornographic class.

In WebGuard, we have chosen the a priori error rates from the cross validation results on MYL learning dataset in order to ensure that the pornographic websites are filtered at maximum. We might also choose other validation results as they differ quite few or global error rates in our formula if we want to a have the best balanced behaviour of our filter both on pornographic and non pornographic classes. As the best a priori rate was a little bit more than 0.03, we set $\delta = 0,03$. After several experiments, we fixed n = 5 giving the best result on MYL test dataset. Using data from Fig. 3, we obtained the relative weight of the five data-mining algorithms as illustrated in Fig. 12, which shows the individual success rates of these algorithms.

As we can see from Fig. 12, the algorithm having the best performance on MYL test dataset, i.e. Sipina 5.22, received the most important weight. The final classification decision of a website is made thus on the basis of the following formula:

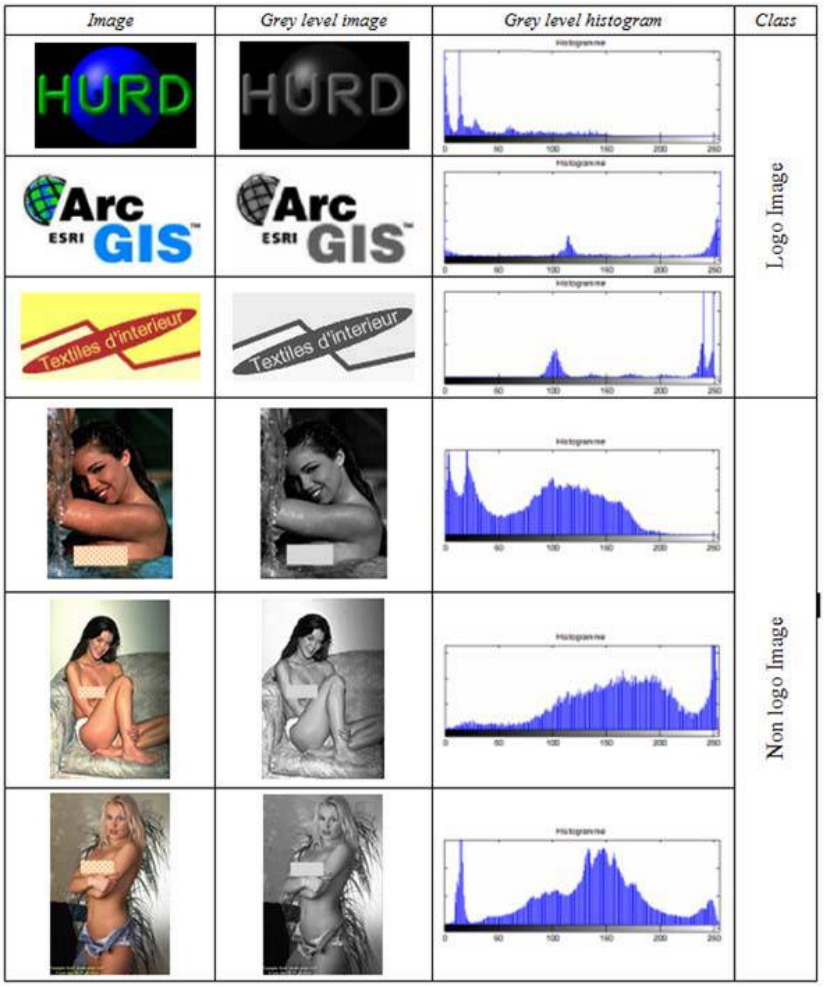

Fig. 11. Logo and non logo image samples and their histograms.

$$\chi = \sum_{i=1,N} \gamma_i \chi_i \geq \tau \text{ with } \chi \text{ ranging from } 0 \text{ to } 1$$

τ is defined as the sensitivity to "objectionable" content. Indeed, if τ is set to 0, WebGuard is the most sensitive as a website is classified as pornographic if just one data-mining algorithm does so. On the contrary, if if τ is set to 1, WebGuard appears as less sensitive as it classifies a website as pornographic only if all the five data-mining algorithms do so. The principle of majority voting used in our experiments for smoothing the classification results corresponds to set $\tau = 0.42$.

IX. Concluding Remarks and Future Work

In this paper, we studied and highlighted the use of skin colour related visual content-based analysis along with textual and structural content-based analysis for improving Web filtering through our machine learning based pornographic website classification engine WebGuard. Using MYL test dataset consisting of 200 adult websites and 200 non pornographic websites, we compared WebGuard-TS using only textual and structural content-based analysis of a webpage with WebGuard-TSV when WebGuard-TS is further cascaded with a skin colour related visual content based classification engine (WebGuard-V). The experimental results first showed that structural content-based analysis, such as hyperlinks, used in WebGuard-TS along with classical textual content-based analysis, has a very good discriminative power [26], [27], [34] as WebGuard-TS already achieved 96.1% classification accuracy rate. Moreover, the visual content-based analysis can further improve the classification performance as WebGuard-TSV scored 97.4% classification accuracy rate when cascading WebGuard-TS and WebGuard-V. The experiment of WebGuard-TSV on a black list of 12 311 adult websites manually collected and classified by the French Ministry of Education further confirms the interest of considering visual content based analysis as WebGuard-TSV scored 95.62% classification accuracy rate compared to 87.82% classification accuracy rate displayed by WebGuard-TS. The basic framework of WebGuard can apply to other categorization problems of websites which combine, as most of them do today, textual and visual content.

We can thus summarize our major contribution by the use of both structural and visual analysis along with classical keyword based textual content analysis for Web document

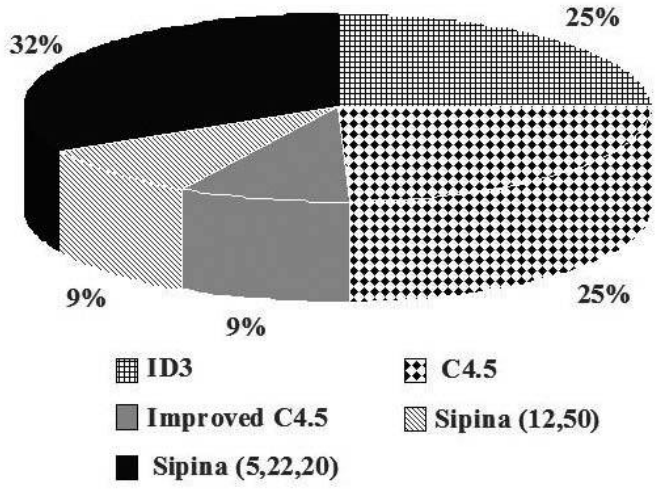

Fig. 12. Relative weights of the five data-mining algorithms in classification decision.

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 [46]. Currently, our dictionary contains more than 300 indicative keywords extracted from six languages. Its elaboration 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 elaboration of dictionary is one of the directions of our future work. Actually, 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 discriminate the best 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. This is a fact that 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.

ACKNOWLEDGEMENTS

Special thanks to anonymous reviewers for their comments which greatly contribute to the improvement of this paper. Our acknowledgement also goes to our students Florian Grus, Nicolas Jouffroy, Audrey Licini, Fabien Perdriel and Phan Le Bao Tuy who carried out the 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*. CampusPress, 2001.
[3] M. Hammami, Y. Chahir, L. Chen and D. Zighed. Détection des régions de couleur de peau dans limage. *Revue RIA-ECA*, 17: 219–231, 2003.
[4] M. Hammami, L. Chen, D. Zighed and Q. Song. *Définition dun modéle de peau et son utilisation pour la classification des images*. Hermés, 2002, pp. 186–197.
[5] M. Hammami, D. Tsishkou and L. Chen. Data-mining based Skin-color Modeling and Applications. *Third International Workshop on Content-Based Multimedia Indexing*, Rennes, France, Septembre 22–24 2003, pp. 157–162.
[6] M. Hammami, Y. Chahir and L. Chen. WebGuard: Web Based Adult Content Detection and Filtering System. *The 2003 IEEE/WIC International Conference on Web Inteligence*, IEEE Computer Society, Halifax, Canada, Octobre 13–17 2003, pp. 574–578.
[7] M. Hammami, D. Tsishkou and L. Chen. Adult Content Web Filtering and Face Detection Using Data-mining based Skin-color Model. *Proceedings of IEEE International Conference on Multimedia and Expo ICME*, June 6–9 2004, Taipei, pp. 403–406.
[8] M. Hammami. Modéle de peau et application á la classification dimages et au filtrage des sites Web. *Thése de doctorat*, Ecole Centrale de Lyon, July 2005.
[9] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1: 81–106, 1986.
[10] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
[11] 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.
[12] Recreational Software Advisory Council on the internet (association that became the Internet Content Rating Association (ICRA) in 1999), www.icra.org.
[13] Cybersitter. 2002 Copyright 1995–2003. Solid Oak Software. www.cybersitter.com.
[14] Net Nanny 4.04. Copyright 2002–2003 BioNet Systems, LLC. www.netnanny.c om.
[15] Norton Internet Security. 2003 Copyright 1995–2003 Symantec Corporation. www.syman tec.com.
[16] Puresight Home 1.6. iCognito Technologies Ltd. www.icognito.com.
[17] Cyber Patrol 5.0. Copyright 2003 SurfControl plc. www.cyberpatrol.
[18] D. A. Zighed and R. Rakotomala. A method for non arborescent induction graphs. Technical report, Laboratory ERIC, University of Lyon2, 1996.
[19] A. Albiol, L. Torres, C. A. Bouman and E. J. Delp. A simple and efficient face detection algorithm for video database applications. *Proceedings of the IEEE International Conference on Image Processing*, Vacouver, Canada, vol. 2, pp. 239-242, September 2000.
[20] H. Wang and S.-F. Chang. A highly efficient system for automatic face region detection in mpeg video. *IEEE Transactions on Circuits and System for Video Technology*,7(4): 615-628, August 1997.
[21] J. G. Wang and E. Sung. Frontal-view face detection and facial feature extraction using color and morphological operators. *Pattern Recognition Letters*, 28(10): 1053-1068, October 1999.
[22] M.-H. Yang and N. Ahuja. Detecting human faces in color images. *Proceedings of the International Conference on Image Processing*, Chicago, IL, October 4–7 1998, pp. 127-130.
[23] M. J. Jones and J. M. Regh. *Statistical Color Models with application to Skin Detection*. Cambridge Research Laboratory, CRL 98/11, 1998.
[24] S. Schüpp, Y. Chahir, A. Elmoataz and L. Chen. Détection et extraction automatique de texte dans une vidéo: une approche par morphologie mathématique. *MediaNet 2002*, Hermés, pp. 73–82 , 2002.
[25] P. Y. Lee, S. C. Hui and A. C. M. Fong. Neural Networks for Web Content Filtering. *IEEE Intelligent Systems*, Sept/Oct"48–57, 2002.
[26] S. Chakrabarti, B. Dom and P. Indyk. Enhanced hypertext categorization using hyperlinks. *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*.
[27] G. W. Flake, K. Tsioutsiouliklis and L. Zhukov. Methods for mining Web communities : bibliometric, spectral, and flow. In: A. Poulovassilis and M. Levene (Eds.), *Web Dynamics*, Springer-Verlag, 2003.
[28] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *WWW7*, Brisbane, Australia, 1998.
[29] J. Cho, H. Garcia-Molina and L. Page. Efficient crawling through URL ordering. *Computer Networks & ISDN Systems*, 30(1–7): 161–172, 1998.

[30] G. W. Flake, S. Lawrence and C. L. Giles. Efficient identification of web communities. *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD-2000)*, Boston, MA, 2000.
[31] Y. Yang, S. Slattery and R. Ghani. A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems*, 2001.
[32] J. Fürnkran. Exploiting structural information for text classification on the WWW. *Intelligent Data Analysis*, 487–498, 1999.
[33] G. Attardi, A. Gulli and F. Sebastiani. Automatic Web page categorization by link and context analysis. In: C. Hutchison and G. Lanzarone (Editors), *Proceedings of THAI-99, 1st European Symposium on Telematics, Hypermedia and Artificial Intelligence*, pp. 105-119, Varese, Italy, 1999.
[34]] E. J. Glover, K. Tsioutsiouliklis, S. Lawrence, D. M. Pennock and G. W. Flake. Using Web structure for classifying and describing Web pages. *WWW2002*, May 7–11 2002, Honolulu, Hawaii.
[35] B. Stayrynkevitch, M. Daoudi, C. Tombelle, H. Zheng *et al*. *Poesia Software architecture definition document*. Technical report, Poesia consortium, December 2002.
[36] E. Karpova, D. Tsishkou and L. Chen. The ECL Skin-color Images from Video (SCIV) Database. *Proceeding of IAPR International Conference on Image and Signal Processing (ICISP2003)*, Agadir, Maroc, 25–27 June pp. 47–52.
[37] Y. Chahir and L. Chen. Efficient Content-based image retrieval based on color homogeneous objects segmentation and their spatial relationship characterization. *Journal of Visual Communication and Image Representation*, 11(1): 302–326, 2000.
[38] W. Mahdi, M. Ardebilian and L. Chen. Text detection and localization within images. PCT/ FR03/ 02406, 31 July 2002.
[39] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.
[40] K. Sato, A. Ohtaguro, M. Nakashima and T. Ito. The Effect of a Website Directory When Employed in Browsing the Results of a Search Engine. *Internationall Journal of Web Information Systems*, 1(1): 43–51, 2005.
[41] S. Y. Chen and X. Liu. Data mining from 1994 to 2004: an application-orientated review. *International Journal of Business Intelligence and Data Mining*, 1(1): 4–21, 2005.
[42] X. Fu and L. Wang. Data dimensionality reduction with application to improving classification performance and explaining concepts of data sets. *International Journal of Business Intelligence and Data Mining*, 1(1): 65–87, 2005.
[43] M. Jeong and D. Lee. Improving Classification Accuracy of Decision Trees for Different Abstraction Levels of Data. *International Journal of Data Warehousing and Mining*, 1(3): 1–14, 2005.
[44] B. Yang and A. R. Hurson. Hierarchical Semantic-Based Index for Ad Hoc Image Retrieval. *Journal of Mobile Multimedia*, 1(3): 235–254, 2005.
[45] J. S.-S. Tang, A. W.-C. Liew and H. Yan. Human Face Animation Based on Video Analysis, with Applications to Mobile Entertainment. *Journal of Mobile Multimedia*, 1(2): 133–148, 2005.
[46] J. D Velásquez, S. Ríos, A. Bassi, H. Yasuda and T. Aoki. Towards the Identification of Keywords in the Web Site Text Content: A Methodological Approach. *International Journal of Web Information Systems*, 1(1): 53–57, 2005.

Mohamed Hammami obtained the PhD in Computer Science from Ecole Centrale at the Lyon Research Center for Images and Intelligent Information Systems (LIRIS) associated to 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 in RNTL-Muse project. He has served on technical conference committees and as reviewer in many international conferences.

Youssef Chahir obtained the M.Sc. degree in Fundamental Computer Science from the University of Technology of Compiegne in 1986, and received, in 2000, a 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 Universit de Nantes in 1984. He obtained the Master degree in 1986 and the PhD in computer science from University of Paris 6. First served as associate professor at Universit de Technologies de Compigne, he joined in 1998 Ecole Centrale de Lyon as Professor where he has been leading an advanced research team on multimedia analysis. Author of more than 90 publications in the multimedia indexing field from 1995, his current research interest includes cross media analysis, multimedia indexing and retrieval, face detection and recognition. He is a member of the IEEE Computer Society.