A Web Pornography Patrol System by Content-based Analysis: In Particular Text and Image

Jantima Polpinij, Chumsak Sibunruang, Somnuk Paungpronpitag, Rapeeporn Chamchong, Anirut Chotthanom

Faculty of Informatics, Mahasarakham University
Mahasarakham 44150 Thailand
{jantima.p, chumsak.s, somnuk.p, rapeeporn.c, anirut.c} @msu.ac.th

***Abstract*— A problem of children being exposed to pornographic web sites on the internet has led to their safety issues. To prevent the children from these inappropriate materials, an effective web filtering system is essential. Content-based web filtering is one of the important techniques to handle and filter inappropriate information on the web. In this paper, we examine a content-based analysis technique to filter the pornographic web sites. Then, our system consists of two primary content-based filtering techniques such as text and image. For text analysis, the Support Vector Machine (SVM) algorithm and N-gram model based on Bayes' theorem is applied and experimented to filter pornographic text for both Thai and English language web sites. Meanwhile, we build and examine an image filtering system with a hierarchical image filtering method. It consists of two main processes such as normalized R/G ratio which is using the pixel ratios (red and green color channels) and human composition matrix (HCM) based on skin detection. The empirical results show that our analysis methods of text and image are more effective for pornographic web filtering. Finally, we have modeled a pornographic web filter using content-based analysis into our Anti-X system.**

***Keywords*—Pornographic web filtering, Content-based analysis, Text filtering, Image filtering**

I. INTRODUCTION

A problem of children being exposed to pornographic web sites on the internet has led to their safety issues. This is because pornographic information can cause children to act out sexually against other children. Furthermore, pornographic exposure can mislead child's sexual attitude and orientation. To prevent the children from these inappropriate materials, an effective web filtering system is essential. Thus, in recent years, many researches have brought significant interest in studying and developing a solution to handle and filter inappropriate information on the webs. Content-based web filtering technique is proposed for handling the pornographic problem. This is because it is well-known that pornographic web sites contain many sexually oriented texts, images, and other information that can also help to filter them.

In this paper, we also propose pornographic web filtering system based-on content in particular text and image. The model of pornographic filtering using text-based is constructed by the Support Vector Machine (SVM) algorithm. Text filetering model primarily focused on Thai and English languages. Meanwhile, we propose a hierarchical image analysis method for image filtering stage. It consists of two main processes: normalized R/G ratio and human composition matrix (HCM) based on skin detection. Finally, the results from text and image analysis are integrated and analyzed them together with the Boolean model. Afterwards, if an URL is analyzed as a pornographic web, it is stored into a blacklist. Finally, these URLs are used for blocking inappropriate material. Our pornographic web filtering system is called *Anti-X system*.

The remainders of the paper are organized as follows. We begin in Section 2 by discussion of background and related work. In Section 3, we describe our research methodology. The experimental design, objectives and experimental results discussed in Section 4. In Section 5, we conclude.

II. LITERATURE REVIEWS

A. Exposure to the Pornographic Websites Problem

The growth in the number of pornographic web sites coupled with the increase in the number of children using the internet leads to a heated debated about safety issues. From several studies (such as [1], [2], and [3]), the psychologists show that pornography can make the children be victims of sexual violence and cause them to act out sexually against other children. Moreover, pornographic exposure can mislead child's sexual attitude and orientation.

According to Donna Rice Hughes's report in [2], there were more than 1.3 million pornographic web sites and 260 million web pages as of September 2003. More than 32 million individuals visited pornographic web sites since September 2003. The cyber-sex industry generates approximately $1 billion annually and is expected to grow to $5-7 billion over the next 5 years. In the year of 2000, "sex" is the number one top searching topic on the Internet. In addition, the report in [4] has also indicated more than 116 thousands child pornography requests per day. 20% of the children surfing the internet were received sexual solicitation. Even the kindergarten students, who are less than 6 years old, have been exposed to the pornographic materials. Exposure to pornography (such as, pornographic materials related to rape or child molestation) threatens to make children victims of sexual violence. Also, the

exposure to pornography often results in sexual illnesses, unplanned pregnancies, and sexual addiction [1]. The Internet has proven to be a useful tool for pedophiles and sexual predators as they distribute child pornography, engage in sexually explicit conversations with children, and seek their victims in chat rooms.

Pornography can cause children to act out sexually against other children. According to [3], Kavanagh has pointed out that children often imitate what they have seen, read, or heard. The previous studies have suggested that exposure to pornography can prompt the kids to act out sexually against younger, smaller, and weaker children. Experts in the field of childhood sexual abuse have also reported that any premature sexual activity in children always suggests two possible stimulants: experience and exposure.

Pornographic exposure can persuade child's attitude into a bad way and mislead child's development. According to Dr. Victor Cline (a psychologist) reported in [3], there are a certain critical periods of childhood where a child's brain is being programmed for sexual orientation. During this period, the mind will develop a feeling for what the person will be attracted to. Being exposed to healthy sexual attitudes during this critical period can result in the child development in a healthy sexual way. Conversely, exposure to deviant pornography during this period can permanently imprint the abnormality of the child sexual orientation.

These bad effects of pornographic exposure have brought the society a serious concern about child safety issues. There have been several proposals to protect them against these pornographic materials such as, using web filtering system.

B. *A Study of Web Filtering Techniques*

Current implementations of web filtering use the techniques of URL blocking (black-list and white-list), keyword blocking, and rating system. Yet, web filtering system based on these three techniques are insufficiently accurate and do not cope well with the ever-changing web. URL blocking uses *a list of URLs* [5] and there are two types of list that can be used: black-list and white-list. The black-list contains URLs that must be blocked. It is a class of inappropriate web sites. The white-list on the other hand, contains a list of permissible sites. The main problem with URL-blocking is that new sites emerge quickly and continually. Therefore, it is difficult to maintain complete and up-to-date lists. Keyword filtering [5, 6] uses *a list of keywords* to identify undesirable web pages. If a page contains a certain number of keywords found in the list, it is considered undesirable. However, the main problem of this method is over-blocking since the meanings of words may depend on the context. For example, a site of sex education may be blocked because of the occurrence of the word "*sex*". In addition, the system is easily defeated using words intentionally or unintentionally misspelled. For example, a malicious site can replace the word "pornographic" with "pornographic" to thwart the filtering system. Such replacement will have little effect on the readability of the page by the user, but it will make keyword filtering significantly more difficult. Therefore, a content-based analysis technique [5] is of interest for the web filtering task. This is because it is well-known that pornographic web sites contain many sexually oriented texts, images, and other information that can also help to filter them.

III. THE RESEARCH ARCHITECHTURE

This section describes the details of each module of our content-based filter. The overall architecture of our research methodology is shown in Figure 1. Our Anti-X system is planned to be installed at the proxy server (web cache) of any school network. It has an important role to automatically detect if the visited web site is pornographic by analyzing the materials (images, texts, and links) of the visited web site. The pornographic web site is then blocked and listed in the blacklist databases. Therefore, it can not be accessed anymore.

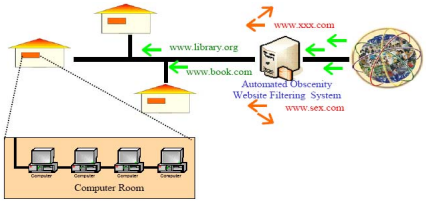


Fig 1. the Anti-X Project

Figure 2 shows the architecture of Anti-X. It is comprised of a proxy system, a URL filtering mechanism and a web content analyzer. There are two main components -- text and image analyzers.

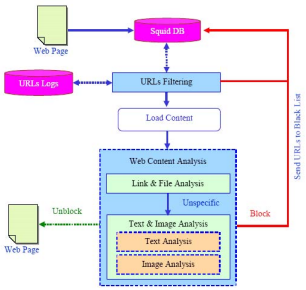


Fig 2. Architecture of the Anti-X system

Each content-based analyzer can be described following.

A. *Text Analysis Method*

In this process, a model of web filtering is built by a supervised machine learning algorithms: Support Vector Machines (SVMs). Then, it is a case of binary text classification, since it involves the classification of incoming documents into two disjoint categories: the pornographic web sites and the non-pornographic web sites. The SVM is a learning method introduced by Vapnik [19] based on the structural risk minimization principle from computation learning theory that finds a maximal margin separating hyperplane between two classes of data. Our text filtering method [7] can be explained as below.

Before learning method, a text collection is transformed into a representation which is suitable for computation. The

ordinary way of document representation is usually as a structured "bag of words" [8]. It will contain each unique word that becomes a feature, including the number of times the word occurs in the document and term weighting that uses the popular term weighting technique (*TF-IDF*) [9]. We also apply the N-gram technique [10] for our work. An N-gram is a sub-sequence of *n* items from a given sequence. An n-gram of size 1 is a "unigram"; size 2 is a "bigram"; size 3 is a "trigram"; and size 4 or more is simply called an "n-gram". N-grams are a popular technique in statistical natural language processing. For parsing, words are modeled such that each N-gram is composed of *n* words. In this work, we apply both unigram and bigrams. Then, bigrams is extracted based on a probabilistic model (Bayes' theorem) following.

$$P(w_i|w_{i-1}) = P(w_{i-1}, w_i)/P(w_{i-1}) \quad (1)$$

Where, the probability of a word w_i given the prior word w_{i-1} is equal to the probability of their bigram, or the co-occurrence of the two words $P(w_{i-1},w_i)$, divided by the probability of the prior word.

In general, pornographic phrases of bigrams often start with these words such as "*adults*", "*sex*", "*free*" etc. Finally, we used these term word as the features. However, before containing in the structured bag of words, the Chi-squared technique (χ^2) [11] is applied to reduce feature size.

For model of text filtering, let $x_1,..,x_i \in \Re^N$, where x_i is a feature vector, be a given training set. The basic concept of SVMs [9] is to build a function that takes the value +1 in a "relevant" region capturing most of the data points, and -1 elsewhere. In addition, let Φ: $\Re^N \rightarrow F$ be a nonlinear mapping that maps the training data from $\Re^N$ to a feature space *F*. Therefore, the dataset can be separated by the following primal optimization problem:

$$\textit{Minimize}: \quad V(\boldsymbol{w}, \xi, \rho) = \frac{\|w\|^2}{2} + \frac{1}{\nu l}\sum_{i=1}^{l} \xi_i - \rho \quad (2)$$

$$\textit{Subject to}: \quad (\mathbf{w} \cdot \Phi(\mathbf{x}_i)) \geq \rho - \xi_i, \ \xi_i \geq 0, \quad (3)$$

where $\nu \in (0,1)$ is a parameter which lets one control the number of support vectors and errors, ξ is a measure of the misclassification errors, and ρ is the margin. When we solve the problem, we can obtain *w* and ρ. Given a new data points $\boldsymbol{x}$ to classify, a label is assigned according to the decision function that can be expressed as follows:

$$f(\mathrm{x}) = sign\ ((w \cdot \Phi(x_i) - \rho) \quad (4)$$

where α_i are Lagrange multipliers and we apply the Kuhn Tucker condition. We can set the derivatives with respect to the primal variables equal zero, and then we can get:

$$\boldsymbol{w} = \Sigma \alpha_i \cdot \Phi(x_i) \quad (5)$$

where α_i are Lagrange multipliers and the Kuhn Tucker condition is applied. We can set the derivatives with respect to the primal variables equal to zero. In addition, there is only a subset of points $\boldsymbol{x}_i$ that lies closest to the hyperplane and has nonzero values α_i. These points are called support vectors. Instead of solving the primal optimization problem directly, the dual optimization problem is given by:

$$\textit{Minimize}: \quad W(\alpha) = \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \quad (6)$$

$$\textit{Subject to}: \quad 0 \leq \alpha_i \leq \frac{1}{\nu l}, \sum_i \alpha_i = 1 \quad (7)$$

where $K(x_i, x_j) = (\Phi(x_i), \Phi(x_j))$ are the kernels functions performing the non-linear mapping into feature space based on dot products between mapped pairs of input points. They allow much more general decision functions when the data are nonlinearly separable and the hyperplane can be represented in a feature space. The kernels frequency used is polynomial kernels $K(\mathrm{x}_i, \mathrm{x}_j) = ((\mathrm{x}_i \cdot \mathrm{x}_j)+1)^d$, Gaussian or RBF (radial-basis function) kernels $K(\mathrm{x}_i, \mathrm{x}_j) = \exp(-\|\mathrm{x}_i - \mathrm{x}_j\|^2/2\sigma^2)$. We can eventually write the decision from equation (3) and (4) and the equation can be illustrated as follow:

$$f(\mathbf{x}) = sign\ (\Sigma \alpha_i K(x_i, x) - \rho) \quad (8)$$

For SVM implementation, we use and modify *LIBSVM* tools from National Taiwan University [3]in our experiments, since we select the RBF kernels for model building. Finally, we tested the pornographic filter by F-measure technique.

B. Image Analysis Method

Our method of image filtering is improved from our previous work [12]. It consists of two sub-processes: normalized R/G ratio [13, 14] and composition based on skin detection [15]. The first process is to use pixel ratios analysis based on the normalized R/G ratio, it used two color channels (red and green) into consideration. Finally, it is to analyze with human composition matrix (HCM) based on human skin detection.

- *Normalized R/G Ratio*

Normalized RGB [13] is a representation, which is easily obtained from the RGB values by a simple normalization procedure:

$$r = \frac{R}{R+G+B} \qquad g = \frac{G}{R+G+B} \qquad b = \frac{B}{R+G+B} \quad (9)$$

The sum of the three normalized components is known ($r+g+b = 1$). As the dependence of *r* and *g* on the brightness of the source, RGB color is diminished by the normalization. However, it was observed that skin invariably contains a significant level of red and green, especially red, and using this observation, certain values of *R*/*G* ratio were used as skin presence indicators [14]. Therefore, we applied only two color tones (red and green) for image consideration in our work, since the *R*/*G* ratio is significant in many skin colors, including those, that are commonly found in African, Asian, and Caucasian skin.

$$r = \frac{R}{R+G+B} \qquad g = \frac{G}{R+G+B} \quad (10)$$

Let $L_{\lim}$ be the lower bound of human skin and $U_{\lim}$ be the upper bound of human skin.

$$\mathrm{L_{lim}} < \mathrm{R/G} < \mathrm{U_{lim}} \quad (11)$$

In this work, we use *L*lim = 1.0559 and *U*lim = 1.8617 that were calculated from prior images (1,000 pornographic images and 1,000 non-pornographic images) by mean value of R and mean value of G. The ratio of these means (R/G) is the *normalized* R/G ratio. Furthermore, the pornographic image data and non-pornographic image data are sorted. We have found that the R/G ratio of a non-pornographic image typically has the normalized R/G ratio as an upper bound, while a pornographic image typically has the normalized R/G ratio as a lower bound.

However, if a web site can not be blocked by the normalized R/G ratio comparison phase, it is passed to the next phase, which is HCM based on skin detection analysis.

- *Human Composition Matrix based on Skin Detection*

Human Composition Matrix (HCM) process is to segment the image into a square matrix, usually having dimensions 6x6 (as shown in Figure 3). The HCM technique will considers each image in the square matrix format.

After an image is in the square matrix format, it is analyzed using human skin detection technique. Human skin detection [15] is done by detecting human skin pixels in an image. The output is a binary image defined on the same pixel grid as the input image.

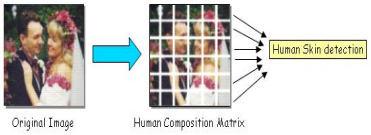


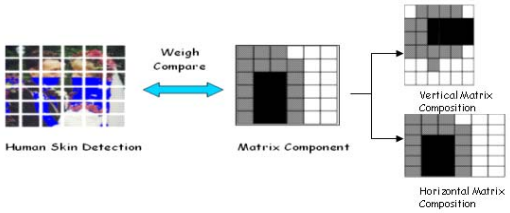
Fig 3. Human Composition Matrixes

A labeled classifier training set is used to construct skin and non-skin color models for each block. Skin and non-skin histogram models are built using the classifier training set of images. Given skin and non-skin histograms, the probability that a given color value belongs to the skin and non-skin classes can be computed using the formula of Jones and Rehg [15] following.

$$P(rgb|skin) = s[rgb]/T_s \quad (12)$$

$$P(rgb|\neg skin) = n[rgb]/T_n \quad (13)$$

where $s[rgb]$ is the pixel count contained in bin rgb of the skin histogram, $n[rgb]$ is the equivalent count from the non-skin histogram, T_s and T_n and are the total counts contained in the skin and non-skin histograms, respectively. After that, a ratio of skin and non-skin histograms is calculated and we use the skin ratio for image analysis. However, we will consider in three classes of human skin: high, medium, and low of skin tones (see in Figure 2). Finally, we tested more matrix sizes such as 3x3, 4x4, 5x5, 6x6, 7x7, 8x8, and 9x9, but the empirical results showed that the 6x6 matrix size is most effective for this task.



Where
The block that contains the highest of skin pixels.
The block that contains the medium of skin pixels.
The block that contains the lowest of skin pixels.

Fig 4. Human Skin Detection

C. *Considering the Both Results of Text and Image by Using Boolean Model*

After the system gives the results of text and image analysis, it is stored these results and represented as a set of results. If a result of each analysis is analyzed as a pornographic website, it turns a result as 1. Meanwhile, if a result is analyzed as a general website, it turns a result as 0. Afterwards, they are considered both results together by using the Boolean query technique [16]. Queries are Boolean expressions of results that are connected by operation 'OR'. Thus, given a Boolean query, the system returns every result that makes the query logically true. It is also called exact match. Then, the final result is relevant or not. It has no partial matches or ranking. We primarily apply this technique for our final analysis because our work is not difficult to express complex requests and control the number of results retrieved.

IV. The Experimental Results

A. *The Experimental Results of Filtering by Text Analysis*

In order to observe characteristics of our system, we do not use any standard dataset. We gathered 1,200 web pages for Thai (400 web sites) and 1,200 web pages for English (400 web sites) in our experiment. We used a collection of pornographic and non-pornographic pages to train a model, which can be used to decide if a given web page is pornographic. We randomly selected 800 pages as a training set and 400 web pages as a test set. After testing by F-measure [17], the results of the model based on unigrams features are shown in

Table 1. The experimental results of the SVM filtering models by using Unigrams

Language	BOWs type	Accuracy (%)	Over-Blocking (%)
English	Unigram	98.00	1.00
Thai	Unigram	95.00	2.00

In Table 1, although the results of the SVM text model are good in terms of accuracy, they show higher over-blocking[1] in

[1] If a general web site is analyzed as a pornographic web site, it actually is not. This problem is called ***over-blocking***.

case of unigram using, especially for grey-zone[2] web pages. In fact, the SVM algorithm is sensitive to error. That means a relatively small number of the unsuitable features can dramatically decrease the performance. Thai web sites may contain content that is too hard to analyze. For example, a Thai word can be used for several cases given several meanings. So, they can lead to unclear analysis. The BOW may contain the unsuitable features. It may lead to poor accuracy of the model. So, if the BOW contains the suitable features, it may lead to better accuracy of the SVM text model.

Afterwards, we apply bigrams to improve the pornographic web filters. The results can be shown in Table 2.

Table 2. The experimental results of the SVM filtering models by using Bigrams

Language	BOWs type	Accuracy (%)	Over-Blocking (%)
English	Bigrams	100.00	0.00
Thai	Bigrams	96.00	1.00

The results in Table 2 show that the accuracies of the SVM text model can be increased not only in the failure rate of accuracy but also the blocking effectiveness. That means the effectiveness of the SVM text model can be increased with a small BOW if the BOW consists of the suitable features. Also, the noun-phrase based on the bigrams can improve the pornographic web filters. However, after the pornographic web filtering is improved by using the both of features, the results can be shown in Table 3.

Table 3. The experimental results of the SVM filtering models by using the both of features.

Language	Accuracy (%)	Over-Blocking (%)
English	100.00	0.00
Thai	95.00	0.00

In final, the SVM text classifier has less over-blocking problem. This may be because the solution of SVM gives an optimal hyperplane, which is a decision boundary between non-pornographic and pornographic. However, the SVM is more sensitive to error. That means, a relatively small number of the unsuitable features can dramatically decrease the performance of SVM. In a Thai web site, it may consist of the contents, which are too hard to analyze. For example, a word can be used for several cases and several meanings. Hence, they lead to unclear analysis. The BOW may contain the unsuitable features.

B. The Experimental Results of Filtering by Image Analysis

In order to observe characteristics of our system, we gathered 4,000 pornographic images and the other 4,000 non-pornographic images for our experiment. We then randomly selected 1,000 pornographic images and 1,000 general images as a training set. We experiment with normalized R/G Ratio and Human Composition Matrix technique. After testing, the experimental results can be presented in Table 4. As the results, this would demonstrate that the HCM can achieve substantial improvements. HCM is based on skin detection model, which is built from a skin pixel-probability or skin pixel count approach. So, it can classify an image to a class according to the skin pixels appearing in the image content, together with the probabilities of the skin pixels learned from a training set. A strong assumption of this approach is that the quantity of interest is governed by the distribution of skin pixel probabilities, and the skin pixels in the image are independent from one another.

Table 4. Testing with 2,022 pornographic images and 1,581 non-pornographic images.

Algorithms	Pornographic Images Effectiveness (%)		Non-pornographic Images Effectiveness (%)	
	Correct	Incorrect	Correct	Incorrect
R/G ratio	90.60	9.40	68.67	31.31
HCM	94.20	5.80	99.67	0.33
Hierarchy	96.50	3.50	99.67	0.33

In addition, we tested the hierarchical image filtering system to filter the web sites on WWW. We used 401 pornographic web sites and 314 non-pornographic web sites for testing. The results are shown in Table 5.

Table 5. Testing with 401 pornographic web sites and 314 non-pornographic web sites by the Hierarchy Image Analysis.

Websites	Effectiveness (%)	Over-blocking (%)
Pornography	93.00	0.00
Non-pornography	96.50	5.00

C. The Experimental Results of Filtering by Text and Image Analysis

Afterwards, the Boolean query technique is applied. The results of text analysis and image analysis are analyzed together by Boolean expressions 'OR'. Thus, given a Boolean query, the system returns every result that makes the query logically true. It is also called exact match. It can be presented in Table 6.

Table 6. The results of Text and Image analysis

Websites	Effectiveness (%)	Over-blocking (%)
Pornography	92.50	0.00
Non-pornography	96.00	0.00

As the results above, the results are always returned based on text analysis because text analysis can give better accuracy than image analysis. Thus, with the Boolean expressions 'OR', the system returns every result that makes the query logically true of text analysis.

V. CONCLUSION AND FUTURE WORK

In this paper, we have built and examined a pornographic web filtering system based on text and image content. Our system is called Anti-X. It consists of main processes. For text

[2] Grey-zone means, a web site often contains the pornographic words in content but it is not a pornographic web site, namely, sex education web site, sex law web site etc.

analyzer, it is built by the Support Vector Machines (SVM) algorithm is applied and experimented to filter the pornographic text for both Thai and English language web sites. Then, we apply N-gram model based on Bayes' theorem to improve the effectiveness of SVM text filtering model. For image analyzer, it consists of two main processes: (i) normalized R/G ratio and (ii) human composition matrix based on skin detection. The first process is normalized R/G ratio, which is the ratio of RED and GREEN from an image. $L_{lim} < R/G < U_{lim}$, where L_{lim} is the lower bound of human skin and U_{lim} is the upper bound of human skin. In this work, we use $L_{lim} = 1.0559$ and $U_{lim} = 1.8617$ that are calculated from prior images. However, if the web site is not blocked by normalized R/G ratio analysis, it will be passed to the next process - HCM. The human composition matrix based on human skin detection will consider each image in the square matrix format. Each image is then separated into 6 x 6 matrixes. Afterwards, the human skin detection technique is applied. We consider in three categories of human skin: high, medium, and low color tones. After testing the system with real-world data, the system was sufficiently accurate. This demonstrates that our proposed techniques can provide more effectiveness for pornographic web filtering based on image content. However, a pornographic web site always contains many sexually oriented texts and images. Therefore, we provide the filtering analysis with both text and image. The results of text analysis and image analysis are analyzed together by the Boolean operation 'OR'. Then, the system always returns every result that makes the query logically true of text analysis.

Next version, we improve the process of HCM. We applied the YcrCb model [18] to provide human skin region because this technique can help to separate chrominance component from luminance component in each human image. Afterwards, it is easier to recognize the differences of kind of human skin such as white, black, yellow, or brown. Based on this, the HCM has an elevated reliability more than the old model. Finally, the HCM is analyzed by using a technique of Neural Network [19].

ACKNOWLEDGMENT

This paper is a part of our *Anti-X* project, which has been supported by the National Electronics and Computer Technology Centre (NECTEC) of Thailand. The project aims at building a system to automatically detect and block the pornographic web sites from school networks of Thailand. Valuable ideas and feedback were received from Dr. Richard Booth. Furthermore, we are also grateful to Mr. Canasai Kruengkrai for a fruitful discussion on Machine learning algorithms. The system is implemented by our research assistants, Mr. Witoon Donprontan and Mr. Supot Chao-nong.

REFERENCES

[1] V. B. Cline. Pornography's Effects on Adults and Children. In *New York: Morality in Media*, vol. 11, 1990.

[2] D. R. Hughes. The Internet safety site of Enough.org. http://www.protectkids.com/, Last accessed: February 2004.

[3] S. J. Kavanagh, Protecting Children in Cyberspace. pp 58-59 Springfield. VA: Behavioral Psychotherapy Center. 1997.

[4] TopTenREVIEWS, "Internet Pornography Statistics", http://internet-filter-review.toptenreviews.com/internet-pornography-statistics.html, Last accessed: February 2007.

[5] P. Y. Lee, S. C. Hui, A. Cheuk & M. Fong. Neural Networks for Web Content Filtering. *IEEE Intelligent Systems*. ISBN 1094-7167. 2002.

[6] R. Du, S. Reihanehi & W. Susilo. *Web Filtering Using Text Classification*. Technical Report. Centre for Communication Security. School of Information Technology and Computer Science. University of Wollongong. Australia. 2002.

[7] J. Polinij, A. Chotthanom, C. Sibunruang, R. Chamchong & S. Puangpronpitag. Content-Based Text Classifiers for Pornographic Web Filtering. *IEEE International Conference on Systems, Man and Cybernetics* (*SMC*). Vol. 2, pp 1481 – 1485. 2006.

[8] R. Baeza-Yates & B. Ribeiro-Neto. Modern Information Retrieval. The ACM Press, New York.1999.

[9] T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *In Proceedings of the 10th European Conference on Machine Learning*, 1997.

[10] T. Pedersen. Lexical Semantic Ambiguity Resolution with Bigram-based Decision Trees. *In Proceedings of Conference on Intelligent Text Processing and Computational Linguistics* (*CICLing*) Mexico. 2001

[11] Y. Yang & J.O. Pederson. A Comparative Study on Features selection in Text Categorization. *Proceedings of the 14th international conference on Machine Learning* (*ICML*). pp 412-420. Nashville, Tennessee. 1997.

[12] C. Sibunruang, J. Polpinij, R. Chamchong, A. Chotthanom & S. Puangpronpitag. Web Pornography Patrol System Based on Hierarchical Image Filtering Techniques. *Proceedings of the 2006 Joint Conference on Information Sciences* (*JCIS*). Taiwan. 2006.

[13] V. Vezhnevets, V. Sazonov & A. Andreeva. A Survey on Pixel-Based Skin Color Detection Techniques. Graphics and Media Laboratory, Faculty of Computational Mathematics and Cybernetics, Moscow State University, Moscow, Russia. 2003.

[14] T. Wark & S. Sridharan. A syntactic approach to automatic lip feature extraction for speaker identification. *In ICASSP*, 3693–3696. 1998.

[15] M. J. Jones & J. M. Rehg. Statistical Color Models with Application to Skin Detection. *International Journal of Computer Vision*. 2001.

[16] R. R. Korfhage. Information Storage and Retrieval. John Wiley & Sons. Canada. 1997.

[17] R. Baeza-Yates & B. Ribeiro-Neto. Modern Information Retrieval. The ACM press. New York. 1999.

[18] L.L. Cao, X.L. Li, N.H. Yu & Z.K. Liu. Naked People Retreival Based on Adaboost Learning. *In Proceeding of the First International Conference on Machine Learning and Cybernetics*. Beijing, China. 2002.

[19] L. Fausett. Fundamentals of Neural Networks Architechture, Algorithms, and Applications. New Jersy: Prentice-Hall International. 1994.