Distinguishing Medical Web Pages from Pornographic Ones: An Efficient Pornography Websites Filtering Method

Jyh-Jian Sheu
(Corresponding author: Jyh-Jian Sheu)
College of Communication, National Chengchi University
Taipei City 11605, Taiwan (R.O.C.)
(Email: jjsheu@nccu.edu.tw)
(Received Dec. 12, 2016; revised and accepted Feb. 19, 2017)

Abstract

In this paper, we apply the uncomplicated decision tree data mining algorithm to find association rules about pornographic and medical web pages. On the basis of these association rules, we propose a systematized method of filtering pornographic websites with the following major superiorities: 1) Check only contexts of web pages without scanning pictures to avoid the low operating efficiency in analyzing photographs. Moreover, the error rate is lowered and the accuracy of filtering is enhanced simultaneously. 2) While filtering the pornographic web pages accurately, the misjudgments of identifying medical web pages as pornographic ones will be reduced effectively. 3) A re-learning mechanism is designed to improve our filtering method incrementally. Therefore, the revision information learned from the misjudged web pages can incrementally give feedback to our method and improve its effectiveness. The experimental results showed that each efficacy assessment indexes reached a satisfactory value. Therefore, we can conclude that the proposed method is possessed of outstanding performance and effectivity.

Keywords: Data Mining; Decision Tree; Medical Web Page; Pornographic Websites Filtering

1 Introduction

Given the anonymity of the Internet, the number of pornographic websites has been increasing steadily in the past decade. The overflow of pornographic information on the Internet has not only imposed impacts on the mental and physical health and values of children or youngsters, but also included by the scholars as one of the causes of physical and mental damage [21]. There are various mechanisms of filtering pornographic websites at present. We study and organize the prevailing filtering methods as the following four categories:

1) Website rating: This method is to filter the web pages by applying rating (or classification) tags [6, 15, 18]. However, this method is lacking in that it is reliant on the initiatives of the website builders. Without any mandatory force, the implementation could not always meet the desired filtering effects.

2) Static filtering: This method works to establish a blacklist of pornographic websites that should be forbidden through website URL, DNS, or the ports of TCP/IP protocols [5, 8, 9]. There are two major types: Site blocking and Internet service blocking. Since the method does not involve the analysis of website contents, there is a high chance that it would make wrong judgments concerning normal websites.

3) Dynamic filtering: It determines whether a website is pornographic or not by analyzing the website content. The analysis on the content and features of website is usually conducted via algorithms, with the aim to discover relevant rules. Dynamic filtering can be divided into two categories: keyword filtering and intelligent content analysis [1, 23].

4) Images filtering: This mechanism would first determine whether what the image represents are human limbs via edge detection and then decide whether the connected groups of limbs could constitute a human figure [3]. In recent years, a lot of filtering methods of pornographic images have been proposed [11, 13, 24, 26]. Moreover, combined with various techniques, numerous intelligent methods of filtering pornographic websites are proposed in succession [4, 7, 12, 27]. However, the excessive computation of scanning images might bring about the low operating efficiency.

Some websites, such as those featuring medical, physical educational and artistic themes, tended to be eas-

ily suspected as phishing websites during the detection process. According to the survey report issued by Pew/Internet, there are about 100,000 medical / health websites around the world, among which over 10,000 are set in the United States [17]. However, among the information on these medical and health care websites, general knowledge on health care, professional information concerning diseases, beauty & slimming, and other health and mental information like sexual knowledge (relevant medicines, methods of birth control, treatment of venereal diseases, etc.); information concerning special periods (pregnancy, parenting, maintenance and physique improvement during puberty); individual mental and health care sharing (fighting pressure, discussions) etc., tend to incorporate pornographic keywords in their contents. For example, the website of American corporative Planned Parenthood Federation[1], this is a legal web page on medical education. But given the existence of suspected pornographic keywords, this web page might be judged as a suspected pornographic web page in spite of its legitimacy.

This study aims to present an efficient mechanism of filtering pornographic websites based on the machine learning technique. In this paper, we apply the uncomplicated decision tree data mining algorithm to find association rules about pornographic web pages. On the basis of these association rules, we propose a systematized method of filtering pornographic websites with the following major superiorities:

1) In order to avoid the low operating efficiency in analyzing photographs, we check only contexts of web pages without scanning pictures. Moreover, the error rates (classify a pornographic website as non-pornographic or a non-pornographic website as pornographic) will be lowered and the accuracy of filtering will be enhanced simultaneously.

2) While filtering the pornographic web pages accurately, the misjudgments of identifying medical web pages as pornographic ones will be reduced effectively.

3) A re-learning mechanism is designed to improve our filtering method incrementally. We apply a supervised machine learning skill to collect any pornographic keywords found newly in the misjudged web pages. Therefore, the revision information learned from the misjudged web pages can incrementally give feedback to our method and improve its effectiveness.

The remainder of this paper is organized as follows. Section 2 introduces the decision tree data mining algorithm. The detailed architecture of our filtering method is shown in Section 3. In Section 4, the experimental results of our method will be presented. Section 5 concludes this paper.

[1] http://www.plannedparenthood.org/

2 Decision Tree Data Mining Algorithm

In this paper, based on decision tree data mining technique, we will propose an efficient systematized method of filtering pornographic websites. The proposed method will analyze the association rules about pornographic web pages and apply them to classify the unknown web pages to be either pornographic or non-pornographic. Decision tree is one of the widely used data mining methods. The technology excels in that it could generate easily understandable association rules and visual features via easy calculations.

There are various decision tree algorithms. Iterative Dichotmiser 3 (called ID3 for short) proposed by Quinlan is one of the most famous and effective decision tree algorithms [19, 20]. According to the study of Stark and Pfeiffer [22], the behavior of ID3 would be better than other improved versions, such as C4.5, CHAID, and CART. Ohmann et al. demonstrated that the quantity of association rules produced by ID3 would not be as numerous as that of C4.5 [16]. Hence, they concluded that ID3 algorithm possessed the superior feature because of the simplicity of rule quantity. Therefore, we adapt ID3 as the data mining technique in this study.

A decision tree is made of a start node (called root node), leaf nodes, and the internal nodes (also called non-leaf node) between the root node and the leaf nodes. In the tree structure, the upper node (called parent node) might branch downward some adjacent nodes (called children nodes). And the final nodes without any branch are called leaf nodes. Suppose that "Target Attribute" is the attribute which is concerned objective of our research. For example, the attribute "web type" ("P" means pornographic websites; "M" means medical websites; "N" means normal websites) is the Target Attribute in this study. Moreover, let "Critical Attributes" be the other important attributes of web pages. The building of the decision tree starts from the root node when all the data instances are contained in the root. The ID3 algorithm will pick out the Critical Attribute with the highest "Information Gain", according to whose values the data instances within the node will be divided into different children nodes. The same process will be repeated by the children nodes on their respective data instances. The algorithm will be ended under two conditions: 1) repeated until all the critical attributes have been selected; or 2) there will be no need of further divisions if the values of the Target Attribute concerning all the data instances within the node are the same. When either of these conditions is met, the current node will be marked as a leaf node. This leaf node will be labeled as the value of Target Attribute possessed by the majority of data instances in this node. Then the algorithm will be stopped and the decision tree is generated completely.

Given a leaf node C, we assume that the value of Target Attribute possessed by the majority of data instances

in C is denoted as $Label(C)$, and $|LabelC|$ is the number of data instances whose Target Attribute's value is the same as $Label(C)$ in C. Then we will compute C's degree of purity (denoted as $Purity(C)$) and degree of support (denoted as $Support(C)$) for this leaf node C. The formulas of $Purity(C)$ and $Support(C)$ are defined as follows:

$$\begin{aligned} Purity(C) &= (|Label(C)|/|C|) \times 100\% \\ Support(C) &= (|C|/N) \times 100\% \end{aligned}$$

where $|C|$ is the number of data instances contained in node C and N is the number of total data instances.

In the resulted decision tree, each path from the root node to a leaf node constitutes an association rule. That's to say, all the internal nodes along the path will serve as the "if" condition for the series of Critical Attributes, together with the "then" outcome represented by the labelled Target Attribute's value of the leaf node, an "if-then" association rule is thus formed.

As compiled by this study, the major calculation steps of ID 3 algorithm are as follows [19]:

1) The algorithm begins from the root node C, when all the data instances are contained in C.

2) If all the data instances within node C have the same value of Target Attribute, then define it as a leaf node, label C by this value, compute $Purity(C)$ and $Support(C)$, and end the algorithm. Otherwise, move on to next step.

3) If all the Critical Attributes have been selected, the values of the Target Attribute concerning the data instances within node C should be examined via majority voting, thus picking out the value boasting the largest number of data instances. Then node C should be defined as a leaf node and labelled by this value, thus computing $Purity(C)$ and $Support(C)$ and ending the algorithm. Otherwise, move on to next step.

4) Calculate the Entropy E(C) for node C through the following expression:

$$E(C) = -\sum_{i=1}^{t} P_i \times \log_2 P_i$$

where t is the number of Target Attribute's values, and P_i = (the number of data instances whose values of the Target Attribute corresponding to the i^{th} value, $1 \le i \le t$, in C)/(the total number of data instances in C).

5) For each Critical Attribute that has not been selected yet (assumed to be attribute α), the Entropy $E^+(\alpha)$ and Information Gain $G(\alpha)$ will be computed by the following expressions respectively:

$$\begin{aligned} E^+(\alpha() &= \sum_{j=1}^{k} (n_j|n) \times E(C_j); \\ G(\alpha) &= E(C) - E^+(\alpha). \end{aligned}$$

In the expressions, we assume that attribute α is supposed to have k values, C_j (for $1 \le j \le k$) represents the subset of the data instances whose values concerning attribute are the same, $E(C_j)$ refers to the Entropy of the subset as calculated through the equation in Step (4), n stands for the total number of data instances within C, and n_j represents the total number of data patterns in the subset C_j.

6) Choose the Critical Attribute that has not been selected yet boasting the highest Information Gain. Assume that the selected Critical Attribute has m values, the children nodes $C_1, C_2, \cdots, C_m$ should be built under this node, to which the data instances of node C should be distributed according to their values of the select Critical Attribute.

7) Respectively treat every children node C_i as node C, $1 \le i \le m$, continue the algorithm recursively from Step (2).

3 An Efficient Pornographic Websites Filtering Mechanism

The objective of this research is to filter pornographic web pages, namely, judging an unknown web page as either pornographic or non-pornographic. While filtering the pornographic web pages, great efforts have been taken to avoid misjudging medical web pages as pornographic ones. For this purpose, medical web pages are set apart from normal web pages in its own category.

In this research, we propose a three-phase systematic method of filtering pornographic websites by applying ID3 decision tree algorithm. The proposed method is possessed of the ability to discriminate between pornographic websites and medical website. Assume that websites will be classified into three categories: "pornographic", "medical", and "normal". Based on the technique of machine learning, our method will discover the association rules about pornographic and medical web pages from training data (known web pages), thus filtering the unknown web pages on the basis of these rules. As illustrated in Figure 1, the structure of the proposed method is comprised of three phases: 1) Training Phase, 2) Classification Phase, and 3) Relearning Phase, which will be introduced as follows.

3.1 The Training Phase

The purpose of this phase is to find association rules of differentiating between pornographic, medical, and normal websites by analyzing training web pages. Then, these association rules will be applied to classify the unknown web pages in the Classification Phase.

In this phase, the training web pages should be examined by the Features Extraction Module to extract their critical features (i.e., the values of Critical Attributes) first. Then, the duplicate copies of training web pages

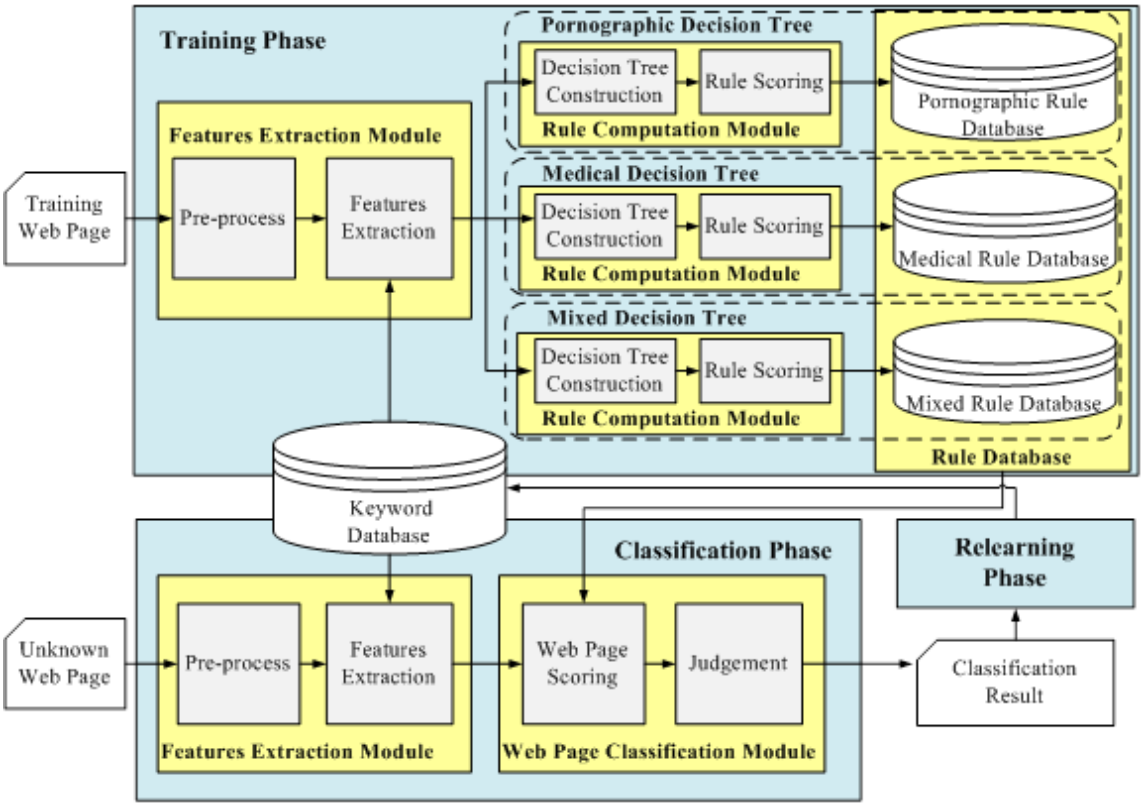

Figure 1: Structure of the proposed method

will be applied simultaneously to construct various decision trees. Note that there are three categories of training web pages: pornographic, medical, and normal. As shown in Figure 1, we construct three decision trees (Pornographic Decision Tree, Medical Decision Tree, and Mixed Decision Tree) and individually equip each decision tree with one Rule Computation Module. By using the copies of pornographic training web pages and normal training web pages as input data, the Rule Computation Module in Pornographic Decision Tree will compute the association rules of distinguishing pornographic websites from normal websites. And the Rule Computation Module in Medical Decision Tree will use the copies of medical training web pages and normal training web pages as input data to compute the association rules of distinguishing medical websites from normal websites. Moreover, the Rule Computation Module in Mixed Decision Tree will use the copies of medical training web pages and pornographic web pages to compute the association rules of distinguishing between medical websites and pornographic websites. Then the resulted rules will be stored respectively into the corresponding rule databases (Pornographic Rule Database, Medical Rule Database, and Mixed Rule Database).

The detailed processes of Features Extraction Module and Rule Computation Module will be discussed as follows.

3.1.1 Features Extraction Module

In the Features Extraction Module, each web page will be analyzed and its Critical Attributes' values will be extracted by applying the following two steps: 1) Pre-process; 2) Features extraction.

The first step is pre-process. In this step, each web page should first be converted into the HTML format, which will be examined by the second step such that its values of Critical Attributes could be verified in the HTML structures.

The second step, features extraction, is designed to discern the critical features of web pages, based on which the suspicious elements of HTML structures that contain relevant keywords will be analyzed. In order to distinguish medical web pages from pornographic ones, judgments will be made based on the features of the HTML head and body, as well as the frequency of medical or pornographic keywords. We study and outline check elements in Table 1 according to the research of Lee et al. [10], which studied and pointed out the parts of this source code mostly likely to be dominated by pornographic keywords. For the sake of convenience, "XXX" will be used to represent the strings that pornographic materials (keywords) might appear. Note that these elements will serve as the Critical Attributes for the computation of association rules in this paper.

As described in Table 1, all these Critical Attributes of each web page will be valued by 0, 1, and 2 by check-

Table 1: The critical attributes used in this study

Type	Critical Attribute		Judgment condition
	No.	Description	
URL	1	Whether there are keywords in the written in HTML Tag of URL.	The URL (URL://XXX) containing pornographic keywords should be set as 1; the URL containing medical keywords should be set as 2; while the URL containing neither should be set as 0.
The head elements	2	Whether there are keywords in the HTML Tag of title.	The HTML Tag ¡title¿XXX¡/title¿ containing pornographic keywords should be set as 1; those containing medical keywords should be set as 2; while those containing neither should be set as 0.
	3	Whether there are keywords in the HTML Tag of link (A).	The HTML Tag ¡link href="XXX"¿ containing pornographic keywords should be set as 1; those containing medical keywords should be set as 2; while those containing neither should be set as 0.
	4	Whether there are keywords in the HTML Tag of link (B).	The HTML Tag ¡link title="XXX"¿ containing pornographic keywords should be set as 1; those containing medical keywords should be set as 2; while those containing neither should be set as 0.
	5	Whether there are keywords in the HTML Tag of metadata (A).	The HTML Tag ¡meta name="author" content="XXX"¿ containing pornographic keywords should be set as 1; those containing medical keywords should be set as 2; while those containing neither should be set as 0.
	6	Whether there are keywords in the HTML Tag of metadata (B).	The HTML Tag ¡meta name="keyword" content="XXX"¿ containing pornographic keywords should be set as 1; those containing medical keywords should be set as 2; while those containing neither should be set as 0.
	7	Whether there are keywords in the HTML Tag of metadata (C).	The HTML Tag ¡meta name="description" content="XXX"¿ containing pornographic keywords should be set as 1; those containing medical keywords should be set as 2; while those containing neither should be set as 0.
	8	Whether there are keywords in the HTML Tag of metadata (D).	Both the HTML Tags ¡meta name="keyword" content="XXX"¿ and ¡meta name="description" content="XXX"¿ containing pornographic keywords should be set as 1; those containing medical keywords should be set as 2; while those containing neither should be set as 0.
The body elements	9	Whether there are keywords in the HTML Tag of hyperlink (A).	The HTML Tag ¡a href="XXX"¿ containing pornographic keywords should be set as 1; those containing medical keywords should be set as 2; while those containing neither should be set as 0.
	10	Whether there are keywords in the HTML Tag of hyperlink (B).	The HTML Tag ¡a¿XXX¡/a¿ containing pornographic keywords should be set as 1; those containing medical keywords should be set as 2; while those containing neither should be set as 0.
	11	Whether there are keywords in the HTML Tag of image (A).	The HTML Tag ¡img src="XXX"¿ containing pornographic keywords should be set as 1; those containing medical keywords should be set as 2; while those containing neither should be set as 0.
	12	Whether there are keywords in the HTML Tag of image (B).	The HTML Tag ¡img alt="XXX"¿ containing pornographic keywords should be set as 1; those containing medical keywords should be set as 2; while those containing neither should be set as 0.
	13	Whether there are keywords in the HTML Tag of image (C).	The HTML Tag ¡img title="XXX"¿ containing pornographic keywords should be set as 1; those containing medical keywords should be set as 2; while those containing neither should be set as 0.
Frequency of keywords	16	There exist 4 to 6 pornographic keywords in the body elements.	The body containing 4 to 6 pornographic keywords should be set as 1; otherwise, as 0.
	17	There exist more than 7 pornographic keywords in the body elements.	The body containing more than 7 pornographic keywords should be set as 1; otherwise, as 0.
	18	There exist 2 to 4 medical keywords in the body elements.	The body containing 2 to 4 medical keywords should be set as 2; otherwise, as 0;
	19	There exist more than 5 medical keywords in the body elements.	The body content containing more than 5 medical keywords should be set as 2; otherwise, as 0;

ing whether the corresponding HTML elements meet the setting conditions.

Note that these Critical Attributes will be examined whether they contain pornographic and medical keywords via the Keyword Database. In this research, the pornographic keywords used are collected from the website SafeSquid[2], and the medical keywords used in this research are collected from the website MedlinePlus [14]. All these pornographic keywords and medical keywords will be stored respectively into Pornographic Keyword Table and Medical Keyword Table of the Keyword Database in advance. Note that the factual category of each training web page is known. If the training web page is pornographic, its Target Attribute should be valued as "P"; if the training web page is medical, its Target Attribute should be valued as "M"; if the training web page is normal, its Target Attribute should be valued as "N". Then, the acquired Critical Attributes and Target Attribute of web pages should be used to build the decision tree in the Decision Tree Construction Module.

3.1.2 Rule Computation Module

As shown in Figure 1, we apply three copies of Rule Computation Module individually to construct three kinds of decision tree and compute their association rules: Pornographic Decision Tree, Medical Decision Tree, and Mixed Decision Tree. This task of the Rule Computation Module contains two steps: 1) Decision tree construction; 2) Rule scoring.

In the first step, Pornographic Decision Tree, Medical Decision Tree, and Mixed Decision Tree will be constructed respectively. The critical characteristics (i.e., Critical Attributes and Target Attribute) of the related training web pages extracted by the Features Extraction Module are set as the input data in each of the three copies of Rule Computation Module. Then ID3 algorithm will be applied to build decision tree and compute the association rule between the Critical Attributes and Target Attribute.

In the second step, we calculate two kinds of score, pornographic score and medical score, for each association rule resulted from the previous step. Each rule will be scored using the formulas based on the values of its degree of support and degree of purity, which are introduced as follows.

Given an association rule R, assume that leaf node of this rule is C, and $Support(C)$, $Purity(C)$, and $Label(C)$ are defined as mentioned earlier. Let $RuleSupport(R)$ be the support degree of rule R with $RuleSupport(R) = Support(C)$. We compute the values of support degree for all rules, and name the maximum one as RS_{MAX} and the minimum one as RS_{MIN}. Let —C— be the number of data instances in the leaf node C. Assume that n_P is the number of data instances concerning the Target Attribute's value is "P" (i.e., pornographic) and n_M is the number of data instances concerning the Target Attribute's value is "M" (i.e., medical) in C. The following three functions are necessary for designing the scoring formula of rules: $PornDegree(R)$, $MedicalDegree(R)$ and $Weight(R)$.

The function $Weight(R)$ calculates the weighted value of rule R by the following formula:

$$Weight(R) = \frac{RuleSupport(R)}{RS_{MAX} + RS_{MIN}} \times 100\%.$$

The function $PornDegree(R)$ implies rule's "intensity" to classify web pages as pornographic, which is defined as follows:

$$PornDegree(R) = Purity(C) \quad \text{if } Label(C) = "P"; \text{ and}$$
$$PornDegree(R) = (\frac{n_P}{|C|}) \times 100\% \quad otherwise.$$

Moreover, the function $MedicalDegree(R)$ implies rule's "intensity" to classify web pages as medical by the following formulas:

$$MedicalDegree(R) = Purity(C) \quad \text{if } Label(C) = "M";$$
$$\text{and } MedicalDegree(R) = (\frac{n_M}{|C|}) \times 100\% \quad otherwise.$$

Finally, we introduce the formulas of computing pornographic score and medical score for rule R respectively: $PornDegree(R)$ and $MedicalDegree(R)$. These two formulas are composed of $Weight(R)$ and either $PornDegree(R)$ or $MedicalDegree(R)$ in a ratio of 3:10, which are described as follows:

$$PornScore(R) = (1 \times PortDegree(R) + 0.3 \times Weight(R)) \times 100;$$
$$MedicalScore(R) = (1 \times MedicalDegree(R) + 0.3 \times Weight(R)) \times 100.$$

By applying the formulas mentioned above, pornographic score and medical score of all rules can be acquired. Then, all rules of three decision trees will be stored into the corresponding rule database, which will be accessed by the Classification Phase to classify unknown web pages.

Moreover, now we define the thresholds in judging the unknown web pages as pornographic or medical for each rule database respectively. In the Pornographic Rule Database, we choose each rule R with $PornDegree(R) \geq 80\%$ and set the minimum pornographic score of the chosen rules as $\lambda(PornRD)$, which will be the threshold of the Pornographic Rule Database for judging the unknown web page is either pornographic or normal used in the Classification Phase. Similarly, we pick out each rule R with $MedicalDegree(R) \geq 80\%$ in the Medical Rule Database and set the minimum medical score of the chosen rules as $\lambda(MedicalRD)$, which will be the threshold of the Medical Rule Database for judging the unknown web page is either pornographic or normal used

[2] http://www.safesquid.com/

in the Classification Phase. Finally, each rule R with $PornDegree(R) \geq 80\%$ in the Mixed Rule Database will be picked out and the minimum pornographic score of the chosen rules will be set as $\lambda(MixedRD)$, which will be the threshold of the Mixed Rule Database for judging the unknown web page is either pornographic or medical used in the Classification Phase.

3.2 The Classification Phase

The purpose of this phase is to examine unknown web pages and classify them as pornographic, medical, or normal. As shown in Figure 1, this phase is comprised of the following two modules: 1) Features Extraction Module and 2) Web Page Classification Module. Firstly, each unknown web page will be inspected by the Features Extraction Module to extract its critical features. Then, the extracted features of this unknown web page will be transmitted to Web Page Classification Module in order to judge its category (pornographic, medical, or normal). The detailed processes of the two modules are described as follows.

3.2.1 Features Extraction Module

The task of Features Extraction Module is basically the same as that of Training Phase. Each unknown web page will be processed by the following two steps: 1) Preprocess; 2) Features extraction. In the first step, each unknown web page will be converted into the HTML format. Then, the second step is to extract the critical features by examining the HTML structure of each unknown web page. By checking the elements outlined in Table 1, the values of 19 Critical Attributes of each unknown web page now can be obtained, which will be used later by the Web Page Classification Module to judge the category of this unknown web page.

3.2.2 Web Page Classification Module

By applying the 19 Critical Attributes extracted in previous module, this Web Page Classification will access the rule databases (Pornographic Rule Database, Medical Rule Database, and Mixed Rule Database) to classify the unknown web pages as pornographic, medical, or normal.

The major steps of algorithm for classifying each unknown web page are as follows:

Step 1. Access the Pornographic Rule Database. This unknown web page will dovetail with some association rule (say, R_1) according to its extracted values of Critical Attributes.

Step 2. Access the Medical Rule Database. Similarly, this unknown web page will dovetail with some association rule (say, R_2) according to the extracted values of Critical Attributes.

Step 3. If $PornDegree(R_1) < \lambda(PornRD)$ and $MedicalDegree(R_2) < \lambda(MedicalRD)$, then this unknown web page will be classified as normal, and stop; else if $PornDegree(R_1) \geq \lambda(PornRD)$ and $MedicalDegree(R_2) < \lambda(MedicalRD)$, then this unknown web page is classified as pornographic, and stop; else if $PornDegree(R_1) < \lambda(PornRD)$ and $MedicalDegree(R_2) \geq \lambda(MedicalRD)$, then this unknown web page is classified as medical, and stop; else if $PornDegree(R_1) \geq \lambda(PornRD)$ and $MedicalDegree(R_2) \geq \lambda(MedicalRD)$, then perform the next step.

Step 4. Access the Mixed Rule Database, and this unknown web page will dovetail with some association rule (say, R_3) according to its extracted values of Critical Attributes. If $PornDegree(R_3) \geq \lambda(MixedRD)$, then this unknown web page will be classified as pornographic; else classify this unknown web page as medical.

3.3 The Relearning Phase

By applying the technique of supervised learning, the task of Relearning Phase is to learn new pornographic or medical keywords incrementally into the Keyword Database. After an unknown web page is judged by Classification Phase, the Relearning Phase will inspect the classification result artificially. In this study, the supervisor will check whether the unknown web page is misjudged. If any misjudgment is produced, the titles and content of the misjudged web pages will then be analyzed and compared to the existing Keyword Database, in order to see whether there are new pornographic keywords or medical keywords. If that is the case, the new keywords will be stored into the Keyword Database.

4 Experimental Design and Results

In this section, we designed and performed experiments to confirm the accuracy and efficiency of the proposed method. In this study, the non-pornographic web pages are composed of medical web pages and normal web pages. In order to measure the performance of this experiment, this study used the decision confusion matrix in Table 2 to estimate the classification results [2]. The purpose of our filtering method is to classify pornographic web pages correctly.

In this research, TP (true positive) means the amount of pornographic web pages that are classified correctly as pornographic; TN (true negative) means the amount of non-pornographic web pages that are classified as non-pornographic web pages. FN and FP refers to misjudgments: FN (false negative) means the amount of pornographic web pages that are misjudged as non-pornographic and FP (false positive) means the amount of

non-pornographic web pages that are misjudged as pornographic.

Table 2: Four cases of judgement

Classification	In reality	
	Pornographic web pages	Non-pornographic web pages
Pornographic web pages	TP (true positive)	FP (false positive)
Non-pornographic web pages	FN (false negative)	TN (true negative)

In this research, the rates of four values of TP, FP, FN, and TN will be computed respectively by the following formulas: $TPR = TP/(TP + FN)$, $FPR = FP/(FP + TN)$, $FNR = FN/(TP + FN)$, and $TNR = TN/(FP + TN)$. Moreover, the three efficacy assessment indexes of "Accuracy", "Precision" and "Recall" will be used to evaluate the filtering accuracy concerning pornographic web pages [23, 25]. Accuracy is used to evaluate the accuracy of the classification results, namely, the proportion of the web pages that are accurately classified to their own categories. It is calculated through the following formula:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Precision is used to evaluate the proportion of pornographic web pages among all the web pages that are judged to be pornographic in nature. It is calculated through the following formula:

$$Precision = \frac{TP}{(TP + FP)}$$

Recall is used to evaluate the proportion of the pornographic web pages that are accurately classified as pornographic, which is calculated through the following formula:

$$Recall = \frac{TP}{(TP + FN)}$$

In this research, "F-measure", which is the harmonic mean of precision and recall, is adopted as one of the measuring indexes of the filtering mechanism. It is calculated through the following formula:

$$Fmeasure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

For example, when the value of precision is too high while the value of recall is too low, it means that although the chances of non-pornographic websites being misjudged are low, the pornographic ones could not be filtered accurately. Under this circumstance, the value of F-measure would be relatively low, thus signifying the poor filtering effects of this method. F-measure is thus a means of evaluation that could combine precision and recall effectively.

The pornographic web pages, medical web pages and normal pages used in this research were compiled from the website urlblackist.com [25]. This website collected all kinds of web pages from various free websites and updates in a continuous manner. This research eliminated inaccurate web pages, web pages without any content, and web pages whose information is not sufficient. Then, we gathered 2250 web pages for the experiments in this study, including 750 pornographic web pages, 750 medical web pages and 750 normal web pages. The numbers of these web pages used in the Training Phase and the Classification Phase of the proposed filtering method were shown in Table 3. Note that the training and unknown web pages should be selected randomly from the three categories.

In the Training Phase, 900 web pages were selected randomly according to the ratio 1:1:1 and trained as three combinations. The training task was performed by three decision trees: Pornographic Decision Tree, Medical Decision Tree, and Mixed Decision Tree. The Pornographic Decision Tree contained 300 pornographic web pages and 300 normal web pages, the Medical Decision Tree contained 300 medical web pages and 300 normal web pages, while the Mixed Decision Tree was the mixture of 300 medical web pages and 300 pornographic web pages.

To confirm the accuracy and efficiency of the proposed filtering method, we performed three experiments, which examined the following performances: (A) the effectiveness of the proposed method in avoiding the misjudgment of medical web pages, (B) the effectiveness of the Relearning Phase, and (C) the stability of the proposed method. These experimental results will be discussed as follows.

(A) The effectiveness of the proposed method in avoiding the misjudgment of medical web pages

The purpose of this experiment was to confirm the effectiveness of the proposed method in avoiding the misjudgment of classifying medical web pages as pornographic ones. In this experiment, we chose randomly 300 medical, 300 pornographic, and 300 normal web pages as the unknown web pages, which would be inputted into the Classification Phase.

Firstly, we performed the Classification Phase without using the Medical Keyword Table (i.e., let the Medical Keyword Table be empty). As shown in Table 4, the number of misjudged medical web pages was 71, while the number of misjudged normal ones was 4. Obviously, the misjudgments of medical web pages were more frequent than that of normal web pages if we omitted the Medical Keyword Table. According to the filtering results of Table 5, the proportion of pornographic web pages that were accurately filtered was (TPR) 97.33%, while the proportion of non-pornographic web pages that were accurately

Table 3: The numbers of web pages used in each phase

	Training web pages in the Training Phase	Unknown web pages in the Classification Phase	Total Total
Pornographic web pages	300	450	750
Medical web pages	300	450	750
Normal web pages	300	450	750
Total	900	1350	2250

Table 4: The classification result of medical web pages and normal ones

	Medical web pages	Normal web pages
The number of misjudged web pages	71	4
The number of web pages judged correctly	239	296
Total	300	300

filtered was (TNR) 87.52%.

Table 5: The efficiency of classification without using the Medical Keyword Table

Indexes	Measurement	Indexes	Measurement
TPR	97.33%	Accuracy	90.79%
TNR	87.52%	Recall	97.04%
FPR	2.67%	Precision	87.52%
FNR	12.48%	F-measure	83.01%

Then, we perform the Classification Phase by applying the Medical Keyword Table. As shown in Table 6, the number of misjudged medical web pages was reduced obviously. This means that after the designed application of the Medical Keyword Table, the filtering accuracy of our method was improved. Moreover, Table 7 recorded the classification efficiency of this experiment. By comparing Table 7 with Table 5, we observed that all the efficacy assessment indexes of Accuracy, Precision, Recall, and F-measure were improved noticeably. Moreover, FPR decreases from 12.48% (before the Medical Keyword Table was used) to 3.67%. Thus, we can deduce that the systematic method proposed in this study will effectively reduce the misjudgments of classifying non-pornographic websites as pornographic ones.

(B) The effectiveness of the Relearning Phase

The purpose of this experiment was to examine the effectiveness of the Relearning Phase of the proposed method. In this experiment, we used 450 medical, 450 pornographic, and 450 normal web pages as the unknown web pages, which would be inputted into the Classification Phase.

The experimental results were shown in Table 8 and 9. The case (I) meant that the Relearning Phase was turned off, and the case (II) indicated that the Relearning Phase was turned on during the classification of unknown web pages. As shown in Table 8, the numbers of misjudged web pages of case (II) were all less than that of case (I), which implied that the Relearning Phase could effectively decrease the probability of misjudgment. Moreover, the values of efficacy assessment indexes were recorded in Table 9. By using the Relearning Phase, the evaluation indicator FPR (the rate of non-pornographic web pages being misjudged as pornographic) decreased from 4.68% to 1.64%. Moreover, the Accuracy increased from 96.21% to 98.26% while the Precision increased from 95.32% to 98.36%. This means that both the Accuracy and Precision were improved after the Relearning Phase was turned on; after the re-learning, TPR (the rate of pornographic web pages being accurately judged as pornographic) increased from 97.95% to 97.99% while FNR (the rate of pornographic web pages being judged as non-pornographic) decreased from 2.05% to 2.01%, showing a slight improvement in terms of the filtering performance concerning pornographic web pages. TNR (the rate of non-pornographic web pages classified accurately as non-pornographic) increased from 95.32% to 98.36%, a significant increase in terms of the classification of normal web pages. FPR decreased from 4.68% to 1.64%, a substantial improvement in terms of the misjudgment rate. These results showed that the relearning mechanism would improve the classification capabilities and performance of the proposed filtering method in this paper.

(C) Testing of the stability

This experiment was set out to investigate whether the classification performance of our method proposed in this paper will be influenced when the data was combined in a different ratio. While the original ratio between normal, pornographic and medical web pages was 1:1:1, some tests were conducted in this experiment over the three kinds of web pages under various ratios, with the aim to guarantee

Table 6: The improved classification result of medical web pages and normal ones

	Medical web pages	Normal web pages
The number of misjudged web pages	18	4
The number of web pages judged correctly	282	296
Total	300	300

Table 7: The classification efficiency of using the Medical Keyword Table

Indexes	Measurement	Indexes	Measurement
TPR	97.33%	Accuracy	96.67%
TNR	96.33%	Recall	97.31%
FPR	2.67%	Precision	96.34%
FNR	3.67%	F-measure	95.86%

Table 8: The number of misjudged web pages

	Medical web pages		Normal web pages		Pornographic web pages	
	Case (I)	Case (II)	Case (I)	Case (II)	Case (I)	Case (II)
The number of misjudged web pages	21	11	8	4	11	7
The number of web pages judged correctly	429	439	442	446	439	443
Total	450	450	450	450	450	450

Table 9: The effectiveness of the relearning phase

Indexes	Measurement		Indexes	Measurement	
	Case (I)	Case (II)		Case (I)	Case (II)
TPR	97.95%	97.99%	Accuracy	96.21%	98.26%
TNR	95.32%	98.36%	Recall	97.89%	98.00%
FNR	2.05%	2.01%	Precision	95.32%	98.36%
FPR	4.68%	1.64%	F-measure	94.06%	98.54%

Table 10: The experimental results of six data groups under various combination ratios

Group No.	Total number of web pages	Ratio	Accuracy (%)	Precision (%)	FNR (%)	FPR (%)
1	900	1:1:2	98.07	97.93	1.78	2.07
		1:2:1	98.22	98.01	1.56	2.00
		2:1:1	98.22	98.65	2.22	1.33
2	900	1:1:3	98.17	97.69	1.33	2.33
		1:3:1	98.06	97.68	1.56	2.33
		3:1:1	98.33	98.66	2.00	1.33
3	900	1:1:5	98.43	97.98	1.11	2.04
		1:5:1	98.39	98.34	1.56	1.67
		5:1:1	98.24	98.69	2.22	1.30
4	600	1:1:2	98.33	98.67	2.00	1.33
		1:2:1	98.33	98.33	1.67	1.67
		2:1:1	98.11	98.22	2.00	1.78
5	600	1:1:3	97.88	97.76	2.00	2.25
		1:3:1	98.33	98.01	1.33	2.00
		3:1:1	98.13	98.25	2.00	1.75
6	600	1:1:5	98.19	98.06	1.67	1.94
		1:5:1	98.50	98.34	1.33	1.67
		5:1:1	98.61	98.88	1.67	1.11

the stability of the filtering mechanism adopted in the current research. Table 10 shows the experimental results of data groups under the different classifications. Note that three tests were conducted for each group, and the three kinds of web pages of each group were combined according the designated ratio. We give an example as follows. Let the total number of web pages in a certain group be 600 and the ratio designated for some test be 1:2:3. Therefore, the web pages in this test will be composed of 100 normal, 200 pornographic, and 300 medical web pages.

Obviously, some changes occurred over the four measuring indicators of Accuracy, Precision, FNR and FPR, though not very substantial changes; when medical web pages accounted for a higher proportion, the FPR (the proportion of non-pornographic pages being misjudged as pornographic) in most groups decreased slightly, but not so much different from the value when the ratio was 1:1:1. This indicated that the filtering results of our method would not be greatly influenced by the changes in the data. In terms of the misjudgment of medical web pages, the values of precision and FPR were fair proof that the method in this research was satisfactory.

5 Conclusions

Concerning the past filtering mechanisms of pornographic web pages, the difficulties in distinguishing medical web pages from pornographic ones have baffled the users of medical websites for a long time. The filtering method proposed in this paper works by selecting the features of the web pages and establishing decision trees according to the category of web pages. Then, the resulted association rules in each decision tree are applied to filter the unknown web pages. To confirm the accuracy and efficiency of the proposed filtering method, we performed three experiments. The first experiment was to examine the effectiveness of the proposed method in avoiding the misjudgment of medical web pages. According to the decrease of FPR, we could deduce that the systematic method proposed in this study would effectively reduce the misjudgments of classifying non-pornographic websites as pornographic ones. The second experiment was to examine the effectiveness of the Relearning Phase. The results showed that the relearning mechanism improved the classification capabilities and performance of the proposed filtering method conspicuously. The experimental results of the third experiment indicated that the filtering results of our method would not be greatly influenced by the changes in the data composition. The Accuracy of this research reached a satisfactory value (greater than 98%). Moreover, the value of F-measure was 98.54%, which showed that the values of Precision and Recall also reached the satisfactory standards, without any figure that's extremely high or extremely low. Therefore, we can conclude that the filtering method proposed in this research is satisfactory because of its outstanding performance and effectivity.

Acknowledgments

This work is supported by the Ministry of Science and Technology, Taiwan, R.O.C. under Grant no. MOST 103-2410-H-004-112.

References

[1] A. Ahmadi, M. Fotouhi, and M. Khaleghi, "Intelligent classification of web pages using contextual and visual features," *Applied Soft Computing*, vol. 11, no. 2, pp. 1638–1647, 2011.

[2] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras, and C. D. Spyropoulos, "An evaluation of naive bayesian anti-spam filtering," in *11th European Conference on Machine Learning*, pp. 9–17, 2000.

[3] M. M. Fleck, D. A. Forsyth, and C. Bregler, "Finding naked people," in *Computer Vision (ECCV'96)*, pp. 593–602, 1996.

[4] M. Hammami, Y. Chahir, and L. Chen, "Webguard: A web filtering engine combining textual, structural, and visual content-based analysis," *IEEE Transactions on Knowledge and Data Engineering*, 1vol. 8, no. 2, pp. 272–284, 2006.

[5] W. H. Ho, and P. Watters, "Statistical and structural approaches to filtering internet pornography," in *IEEE International Conference on Systems, Man and Cybernetics*, vol. 5, pp. 4792–4798, 2004.

[6] ICRA, *Internet Content Rating Association*, Mar. 29, 2017. (`http://www.fosi.org/icra/`)

[7] T. Kajiyama, and I. Echizen, "An educational system to help students assess website features and identify high-risk websites," *Interactive Technology and Smart Education*, vol. 12, no.1, pp. 14–30, 2015.

[8] M. Kanuga, and W. D. Rosenfeld, "Adolescent sexuality and the internet: the good, the bad, and the URL," *Journal of Pediatric and Adolescent Gynecology*, vol. 17, no. 2, pp. 117–124, 2004.

[9] L. H. Lee, and C. J. Luh, "Generation of pornographic blacklist and its incremental update using an inverse chi-square based method," *Information Processing & Management*, vol. 44, no. 5, pp. 1698–1706, 2008.

[10] P. Y. Lee, S. C. Hui, and A. C. M. Fong, "An intelligent categorization engine for bilingual web content filtering," *IEEE Transactions on Multimedia*, vol. 7, no. 6, pp. 1183–1190, 2005.

[11] D. Li, N. Li, J. Wang, and T. Zhu, "Pornographic images recognition based on spatial pyramid partition and multi-instance ensemble learning," *Knowledge-Based Systems*, vol. 84, pp. 214–223, 2015.

[12] T. M. Mahmoud, T. Abd-El-Hafeez, and A. Omar, "An Efficient System for Blocking Pornography Websites," in *Computer Vision and Image Processing in Intelligent Systems and Multimedia Technologies*, IGI Global, pp. 161–176, 2014.

[13] J. A. Marcial-Basilio, G. Aguilar-Torres, G. Sanchez-Perez, L. K. Toscano-Medina, and H. M. Perez-Meana, "Detection of pornographic digital images," *International Journal of Computers*, vol. 5, no. 2, pp. 298–305, 2011.
[14] MedlinePlus, http://www.nlm.nih.gov/medlineplus/.
[15] M. G. Noll, and C. Meinel, "Web page classification: An exploratory study of the usage of Internet content rating systems," in *LIASIT-Luxembourg International Advanced Studies in Information Technologies*, Luxembourg, 2005.
[16] C. Ohmann, V. Moustakis, Q. Yang, K. Lang, and Acute Abdominal Pain Study Group, "Evaluation of automatic knowledge acquisition techniques in the diagnosis of acute abdominal pain," *Artificial Intelligence in Medicine*, vol. 8, no. 1, pp. 23–36, 1996.
[17] Pew Internet, *What the Public Knows About Cybersecurity*, Mar. 22, 2017. (`http://pewinternet.org/`)
[18] PICS, *Platform for Internet Content Selection*, Mar. 29, 2017. (`http://www.w3.org/PICS`)
[19] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.
[20] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Elsevier, 2014.
[21] B. H. Schell and C. Martin, *Cybercrime: A Reference Handbook*, ABC-CLIO, 2004.
[22] K. D. Stark and D. U. Pfeiffer, "The application of non-parametric techniques to solve classification problems in complex data sets in veterinary epidemiology-an example," *Intelligent Data Analysis*, vol. 3, no. 1, pp. 23–35, 1999.
[23] G. Y. Su, J. H. Li, Y. H. Ma, and S. H. Li, "Improving the precision of the keyword-matching pornographic text filtering method using a hybrid model," *Journal of Zhejiang University Science*, vol. 5, no. 9, pp. 1106–1113, 2004.
[24] L. Sui, J. Zhang, L. Zhuo, and Y. C. Yang, "Research on pornographic images recognition method based on visual words in a compressed domain," *IET Image Processing*, vol. 6, no. 1, pp. 87–93, 2012.
[25] URL blacklist service, Mar. 29, 2017. (`http://urlblacklist.com/`)
[26] J. Zhang, L. Sui, L. Zhuo, Z. Li, and Y. Yang, "An approach of bag-of-words based on visual attention model for pornographic images recognition in compressed domain," *Neurocomputing*, vol. 110, pp. 145–152, 2013.
[27] L. Zhuo, J. Zhang, Y. Zhao, and S. Zhao, "Compressed domain based pornographic image recognition using multi-cost sensitive decision trees," *Signal Processing*, vol. 93, No. 8, pp. 2126–2139, 2013.

Biography

Jyh-Jian Sheu is currently an associate professor in College of Communication, National Chengchi University, Taiwan. He received his B.B.A. degree in Management Information Systems from National Chengchi University, Taiwan, and his M.S. and Ph.D. degrees in Computer and Information Science from National Chiao Tung University, Taiwan. His primary research interests include data mining, Internet security, and Big Data.