Learning Association Rules and Tracking the Changing Concepts on Webpages: An Effective Pornographic Websites Filtering Approach

Jyh-Jian Sheu

Abstract—We applied the decision tree algorithm to learn association rules between webpage's category (pornographic or normal) and the critical features. Based on these rules, we proposed an efficient method of filtering pornographic webpages with the following major advantages: 1) a weighted window-based technique was proposed to estimate for the condition of concept drift for the keywords found recently in pornographic webpages; 2) checking only contexts of webpages without scanning pictures; 3) an incremental learning mechanism was designed to incrementally update the pornographic keyword database.

Index Terms—Concept drift, data mining, decision tree, pornographic websites filtering.

1. Introduction

Due to the development of the Internet, the population age of Internet users gradually declined. A social survey conducted in June 2010 in UK shows that about 1/3 children have browsed pornographic websites before the age of 10. Browsing pornographic websites at an early age can easily produce an adverse impact on children; many European countries have formulated relevant laws and regulations to filter pornographic websites[1]. For example, the Netherlands government passed a law to empower the police with relevant rights to filter pornographic websites. However, when strengthening the filtering of pornographic websites, it can often result in the problem of over-blocking. Thus, using appropriate filtering software and technology has become a major issue at present.

Currently, pornographic information spread on the Internet is in the forms of text, pictures, and videos. The pornographic information will import sensational senses into the minds of young people. At the same time, stimulative and provocative words, pictures, and videos of pornographic materials may result in the abnormal gender concepts in the period of character and personality cultivation period. More seriously, it may lead to severe sexual harassment, underage rape, abortion, infectious diseases, and other social problems. As a result, the issue has attracted the attention of families and governments around the world[4,5].

Online pornographic information can be disseminated by any computer in the world. Governments around the world should classify and manage Web content. World Wide Web Consortium (W3C) proposed the Platform for Internet Content Selection (PICS) as a network classification standards system (http://www.w3.org/PICS/). There are many studies on filtering software and a number of commercialized databases of blacklists in the market[4,5]. Moreover, there are some inappropriate content blocking and filtering software for downloading by users. However, in the face of rapid changes in the Internet, how to effectively update and improve filtering accuracy has become an issue worthy of discussion.

In recent years, as data mining technology advances, scholars have applied various machine learning methods to successfully filter pornographic webpages, and such machine learning methods include: Naïve Bayes[6], support vector machines (SVM)[7], and decision tree[8]; methods using natural language as the filter include neural networks[9], K-nearest neighbors[10], the currently common webpage filtering and analysis methods include keyword matching[11], uniform resource locator blocking (URL blocking)[12,13] and image analysis[14-17].

However, these pornographic webpage filtering methods proposed in the past have inadequacies. In keyword filtering, with given keyword database, keywords of higher frequency are often stored in the keyword database. The accuracy of filtering by using keyword only is challenged and doubted since the main problem of keyword filtering is the determination of semantics and times of occurrences. The URL analysis is based on sources and descriptions of URLs. The URL of a pornographic website might contain the fragment word of “SEX” or “PORN”. However, many pornographic websites’ URLs are often composed of irrelevant words or random numbers. Therefore, the effect of filtering by using website URL is quite limited. Most the current URL-related studies propose to establish the whitelists and blacklists, including the blocking of the URLs on the lists. URL on the white list will be allowed. The disadvantage of such a method is that it takes a lot of
costs to maintain and update the URL list to keep the accuracy rate of the filtering system\textsuperscript{18}. The webpage image analysis is to determine a picture as pornographic or not by the color distribution, shape, and vector graphs of the picture. The problem of the filtering method is the lack of accuracy, and it will consume a lot of time. The independent hypothesis of Naive Bayes is too simple to be consistent with the real environment\textsuperscript{19}. Among the data mining machine learning methods, the problem of decision tree techniques is that the growing process of the decision tree is not applicable for streaming data. To address this problem, scholars use the data mining method of concept drift\textsuperscript{20}. By using such a skill, even the content of the data change over time, the decision tree can strengthen learning to solve the problem mentioned above.

The real-world database is not static but changing with time. For example: After the American popular singer Lady Gaga staging a spicy performance in MV, in the same period of time, many articles on the pornographic website may also have a large number of keywords relating to Lady Gaga. After a period of time, as the news event fades, the occurrence of such keywords may be reduced relatively. Similar themes may be used as the eye-catching titles of the pornographic website. The phenomenon of pornographic website content changing with time and the real event is known as “concept drift”\textsuperscript{21,22}. Therefore, the model originally established by using machine learning filter should be updated regularly to fit the real world information.

Solutions to the concept drift problem can be roughly divided into three types as suggested in literature:

1) Instance selection: The main purpose of this method is to find out the data instances relating to current concept. The most common technology is window-based method\textsuperscript{19,20,22}. The data instances of the latest concepts are stored in the window with limited capacity. When data instances of the new concept are added and the window is filled up, the data instances of the old concept will be moved out of the window. The data instances of the current concept will be recorded by using such a method. Moreover, the occurrence of concept drift in the future can be predicted by observing the data instances of the concept in the window. However, it is unfair to control concept drift by the quantity of data instances in the window and it is unable to present the time weights of the data instances in the window. Therefore, scholars proposed new algorithms to dynamically adjust the quantity of windows in order to judge the phenomenon of concept drift effectively\textsuperscript{23}.

2) Instance weighting: This method gives different weights to data instances. Each data instance will be assigned a weight according to its arrival time\textsuperscript{24}. For example, the weight of each data instance is determined by factors including the time of life. Then, the classifier will compute the classification results of data instances in accordance with their weights. Note that a data instance should be disposed if it exceeds the predetermined time scope.

3) Ensemble learning method: This method applies more than two classifiers’ weighted voting to determine the adoption of the classifier. Different classifiers are used to predict the classification results and assign different weights in case of various circumstances. It combines different classifiers and selects out optimal combinations to achieve the best classification results in order to solve the concept drift problem\textsuperscript{25}.

In this paper, we apply the uncomplicated machine learning manner and the decision tree data mining technology to propose an efficient method of filtering pornographic webpages, which is possessed of the capability of tracking the concept drift on webpages. The ultimate goal of this study is to improve the filtering accuracy and reduce the cost of filtering pornographic webpage; through data mining, the proposed method can filter not only URL, but also scan the content (body). Regarding URL features and content features, we use the decision tree algorithm to learn association rules between webpage’s category (pornographic webpage, normal webpage) and these feature attributes. Based on the resulted rules, we will propose a more efficient filtering mechanism of pornographic webpages with the following major advantages:

a) Regarding the solution to the problem of tracking concept drift in pornographic webpages, we propose a weighted window-based technique to estimate for the condition of concept drift for the keywords found recently in pornographic webpages. This technique will help our filtering method in recognizing the pornographic webpages.

b) We check only contexts of webpages without scanning pictures in order to avoid the low operating efficiency in analyzing photographs. Furthermore, the error rates that classify pornographic webpages as normal or the normal ones as pornographic will be lowered and the accuracy of filtering will be enhanced concurrently.

c) An incremental learning mechanism is proposed to increase the pornographic keyword database. By increasing and updating the pornographic keyword database incrementally, it will make the proposed filtering method better suitable for the dynamic environment.

The remainder of this paper is organized as follows. Section 2 introduces the features of pornographic websites and the decision tree data mining algorithm. In Section 3, we describes the detailed architecture of our filtering method proposed in this paper. The experimental results of our method are presented in Section 4. Section 5 concludes this paper.

2. Features of Pornographic Webpage and Decision Tree Algorithm

2.1 Features of Pornographic Webpages

The hypertext markup language (HTML) is mainly used to structure information (such as titles, paragraphs, tables, etc.). It could also describe the appearance and semantics of
text. On the Internet, there are websites that incorporate all kinds of media, such as text, images, videos, and audios, and present them to the users via browsers. This is further achieved through HTML, the markup language that enables the browsers to process and present the contents of raw files. HTML will classify the various features into different elements and attach special marks to signify the composition of these files. The marks that indicate the start and end of the content, as well as to which element the content belongs, are called “Tags”. HTML is thus comprised of various elements, Tags, and attributes. During the operations of HTML, only Tags will be recognized, while paragraphs and capitalization are negligible. Note that each Tag is wrapped up by the paired “<” and “>”. The HTML structure is basically expressed by Tags as follows:

```html
<Element_Name Attribute_Name= "Value"> Content of the Element </Element_Name>
```

The symbol “<” represents the start of this Tag; “Element_Name” and “Attribute_Name” are comprised of English letters; “Value” is comprised of English letters or numbers; the symbol “>” represents the end of the Tag; the symbol “/>” represents the end of symmetrical Tag. Table 1 presents some examples of elements in webpage files.

Fig. 1 is the omitted version of the source code of a real pornographic webpage. The source code is comprised of a series of HTML syntax, and the content is made up of such Tags as the title, content, image, and hyperlink. Lee et al. studied and pointed out the parts of this source code most likely to be dominated by pornographic keywords. Take Fig. 1 for example, a suspected pornographic keyword “Adult” is discovered in the fourth line around the Tag “<title>” and such a statement as “Warning, this part of MostPlays contains adult related content!” is seen on the fourteenth line around the Tag “<strong>”, a warning stating that this is an adult webpage.

Suspected pornographic words are often found on pornographic webpages, such as indecent, provocative, or sex-related text, which all will be called “pornographic keywords” here. The following parts are common places on HTML codes where pornographic keywords are likely to exist. For the sake of convenience, “PORN” will be used to represent pornographic materials (keywords).

### Table 1: Some elements of webpage

<table>
<thead>
<tr>
<th>Elements</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>Display the textual content of the webpage.</td>
<td><code>&lt;body&gt;This is my first paragraph.&lt;/body&gt;</code></td>
</tr>
<tr>
<td>Image</td>
<td>Indicate the type and name of web page images: <code>src</code> refers to the source of the specified image, and <code>alt</code> refers to the texts designed to replace the specified image.</td>
<td><code>&lt;img src= &quot;boat.gif&quot; alt=&quot;Big Boat&quot; /&gt;</code></td>
</tr>
<tr>
<td>Hyperlink</td>
<td>Indicate a hyperlink that will be connected to an external webpage. And a textual description is followed.</td>
<td><code>&lt;a href=&quot;http://www.w3schools.com&quot;&gt;This is a link&lt;/a&gt;</code></td>
</tr>
<tr>
<td>Title</td>
<td>Indicate the name of the webpage title. Displayed on the title bar of the browsers.</td>
<td><code>&lt;title&gt;The XXX adult page&lt;/title&gt;</code></td>
</tr>
<tr>
<td>Metadata</td>
<td>Describe the metadata of webpage, including several kinds of parameters such as “author”, “description”, “keyword”, etc.</td>
<td><code>&lt;meta name=&quot;keyword&quot; content=&quot;family&quot; /&gt; &lt;meta name=&quot;author&quot; content=&quot;Smith&quot;&gt; &lt;meta name=&quot;description&quot; content=&quot;The family of Smith&quot;&gt;</code></td>
</tr>
</tbody>
</table>

Fig. 1. Omitted version of the source code of a real pornographic webpage file.
1) The content of “URL”: Suspected pornographic keywords might be contained in the URLs of pornographic webpages. But since the names of most webpages are expressed in English and meaningless alphabetic arrangement or numbers tend to be used to avoid attention, it is not efficient enough to decide whether a website is pornographic in nature by relying on URLs alone. The ULRs with pornographic keywords might be written in HTML as

\[ \text{URL://www.PORN.com.} \]

2) The page title defined by Tag “<title>”: The defined title will be displayed on the title bar of the browser, and usually the default title in the users’ “favorite” or “bookmark”. It is usually used by pornographic webpages to denote the eye-catching title with pornographic keywords and written in HTML as

\[ <\text{title}>\text{PORN}<\text{title}>. \]

3) The metadata defined by Tag “<meta>”: It describes the metadata of a webpage and includes several kinds of parameters. For example, “author” is used to indicate the author of webpage, “description” is used to record the synopsis of webpage, and “keyword” is used to define keywords of webpage to facilitate the search engines. However, such information will not be displayed while the users are browsing webpages. The pornographic keywords are likely to be written in HTML as

\[ <\text{meta name}="\text{author}\text{ content}="\text{PORN}"/>\]
\[ <\text{meta name}="\text{description}\text{ content}="\text{PORN}"/>\]
\[ <\text{meta name}="\text{keyword}\text{ content}="\text{PORN}"/>. \]

4) The external files or resources denoted by the Tag “<link>”: This Tag is used to set the relationships between the files, namely the hyperlinks. The pornographic materials are likely to be written in HTML as

\[ <\text{link href}="\text{PORN}">. \]

5) The image denoted by the Tag “<img>”: The Tag “<img>” is designed to control the images in webpages. The parameter “src” refers to the source URL of the specified image, while “alt” refers to the text designed to replace the specified image. When the browser of the user does not support the file’s type of this image, these words defined by “alt” will appear instead. The function of “title” is similar to that of “alt”, namely, when the browser does not support “alt”, “title” will be used instead. They are written in HTML as

\[ <\text{img src}="\text{PORN}" \text{alt}="\text{PORN}" \text{title}="\text{PORN}"/>. \]

6) The hyperlink defined by Tag “<a>”: This Tag indicates a hyperlink and a textual description is followed. Both of them are usually utilized by the pornographic webpages as follows

\[ <\text{a href}="\text{PORN}">\text{PORN}</a>. \]

7) The text defined by the Tag “<body>”: This part refers to the textual content of the webpage files, in which the suspected pornographic keywords might be found directly. They are written in HTML as

\[ \text{<body> PORN </body>}. \]

The pornographic webpages may not contain all the above items but parts of the above items. Therefore, by analyzing the pornographic webpage’s features and studying the common places on HTML codes where pornographic keywords are likely to be found, we can design some important attributes to help filter pornographic webpages effectively, which will be discussed in Section 3.

2.2 Decision Tree Data Mining Algorithm

In this paper, we propose an efficient systematized filtering method based on decision tree data mining technique to filter pornographic webpages. Decision tree algorithms are effective data mining methods upon the tree data structure, and they can find the potential association rules of data to predict (or classify) the unknown data whereas the statistical methods can only compute the superficial distribution of data.

Among various decision tree algorithms, Iterative Dichotomiser 3 (called ID3 for short) proposed by Quinlan is the most well-known one[26]. ID3 only supports categorical attributes. However, C4.5 (an improved method of ID3) can support both categorical attributes and numerical attributes simultaneously[27]. Since the data attributes adopted in this paper are the mixture of numerical type and categorical type, we select C4.5 as the major data mining method to build an efficient systematized method of filtering pornographic webpages.

We assume that “target attribute” is the attribute which is concerned objective in this paper (that is, the attribute “webpage’s type”, whose value “P” means it is a pornographic webpage and value “L” means it is a legitimate one) and “critical attributes” are the other influential attributes for classifying pornographic webpages. The process of building decision tree will start from the root node, and all data instances are initially contained in this root node. The C4.5 algorithm will select a critical attribute with the maximum “gain ratio” but not yet selected (the detailed process of computing gain ratio will be described later). Then, all data instances in the root node will be divided into disjoint children nodes according to their values of this selected critical attribute. Subsequently, the same process will be repeated respectively by each of those children nodes for its own data instances[27].

Now we introduce the process of computing the gain ratio. Assume that the current node is C. Given a certain critical attribute A, the gain ratio of A will be denoted as GainRatio(A), which concerns its “information gain” of A (denoted influential Gain(A)) and “entropy” of node C (denoted as E(C)). E(C) is computed by

\[ E(C) = - \sum_{i=1}^{t} \frac{1}{n} \log_2 \left( \frac{p_i}{n} \right) \]

where \( p_i \) is the number of data instances corresponding to the \( i \)th value of target attribute in \( C \), \( n \) is the total number of data instances in \( C \), and \( t \) is the number of target attribute’s values. Then, Gain(A) is calculated by:

\[ \text{Gain Ratio} = \frac{\text{Gain}_A}{\text{Entropy}(C)} \]
Gain(A) = E(C) - E^(+)(A)
E^(+)(A) = \sum_{j=1}^{k} (n_j/n)E(C_j)

where k is the number of values of A, C_j with 1 \leq j \leq k is a subset of C including the data instances corresponding to the jth value of A, and n_j is the number of data instances contained in C_j. Finally, we can compute GainRatio(A) by the following formulas:

GainRatio(A) = Gain(A)/SplitInfo(A)
SplitInfo(A) = \sum_{j=1}^{k} (n_j/n)log_2(n_j/n).

There are two basic stop conditions to end the construction process of the decision tree of the C4.5 algorithm: 1) all critical attributes are selected; 2) target attribute’s values of all data instances in this current node are exactly the same. If any of the two stop conditions is satisfied, the construction process of decision tree will be stopped and this current node will be signified as a leaf node by performing the following ending process.

Assume that C is the current node that will be signified as a leaf. Then, C will be labeled by the value of the target attribute possessed by the majority of data instances in C, which is denoted as Label(C). Let Label(C) be the number of data instances whose target attribute’s value is Label(C) in the node C. Then, compute the degree of purity (denoted as Purity(C)) and degree of support (denoted as Support(C)) for the node C, and end the construction process of the decision tree. The formulas of Purity(C) and Support(C) are defined as follows:

Purity(C) = ([Label(C)]/|C|) \times 100%
Support(C) = (|C|/N) \times 100%

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

In this paper, we propose another stop condition to avoid excessive growth of decision tree as follows. Suppose that the current node is C, if Purity(C) is lower than 20% or higher than 90%, or Support(C) is below 10%, the construction process of decision tree will be stopped; on the contrary, if Purity(C) is between 20% to 90% or Support(C) is higher than 10%, it will go on growing. Now we introduce the detailed procedure of the C4.5 algorithm[27] as follows.

Step 0. All data instances are stored in the root node, select a critical attribute, say A, with the maximum gain ratio. Divide all data instances into disjoint children nodes according to their values of the selected critical attribute A. Then the following steps will be executed for each children node.

Step 1. If the target attribute’s values of all data instances in the current node are exactly the same, then signify this node as a leaf node, perform the ending process, and stop.

Step 2. If all critical attributes have been selected, then signify the current node as a leaf node, perform the ending process, and stop.

Step 3. Treat the current node as node C. If any of Purity(C)<20%, Purity(C)>90%, and Support(C)<10% is satisfied, then signify this node as a leaf node, perform the ending process, and stop.

Step 4. Calculate the gain ratio for each critical attribute that is not selected yet. And select the critical attribute with the maximum gain ratio. Then divide all data instances into disjoint children nodes according to their values of the selected critical attribute.

Step 5. Repeat Step 1 to Step 5 for each of the children nodes generated in Step 4 respectively.

Finally, the decision tree will be built. In the resulted decision tree, each leaf node will be labeled as a value of the target attribute. Each path from the root node to a leaf node will form an association rule such that all internal nodes on this path constructed a row of “if” judgment of the critical attributes. With the “then” result denoted by the labeled value of the leaf node, an association rule of “if-then” pattern will be constructed.

3. Efficient Method of Filtering Pornographic Websites

In this paper, we use machine learning manner and decision tree data mining algorithm to propose an efficient pornographic websites filtering method, which will be possessed of the capability of tracking the concept drift on webpages. We choose the C4.5 algorithm as the major decision tree data mining method applied in this study. Based on the technique of machine learning, the proposed method will learn the association rules about pornographic webpages from training data (known webpages, which have been classified as either pornographic or legitimate one), thus filtering the unknown webpages on the basis of these rules.

As illustrated in Fig. 2, the architecture of the proposed method can be divided into two stages: 1) training stage and 2) classification stage, which will be introduced as follows.

3.1 Training Stage

The purpose of this stage is to find association rules about pornographic websites by analyzing “training webpages”. The training webpages are known webpages that have been classified as either pornographic or legitimate one. Then, the resulted association rules will be applied to classify the unknown webpages in the classification stage.

In the first step of this stage, the training webpages should be examined by the features extraction module to extract their important features which can be used to identify pornographic webpages. Note that the factual category of each training webpage will be regarded as the target attribute in this study. The target attribute of a training webpage will be valued as “P” if it is pornographic,
and valued as “L” if it is legitimate. Moreover, we have designed 16 important attributes as the critical attributes, whose values will be computed by the features extraction module and concept drift scoring module. Then, the decision tree data mining module will perform the C4.5 decision tree algorithm to analyze the training webpages and compute association rules between 16 critical attributes and the target attribute. Finally, we will score the association rules, and store them into the rule database, which will be accessed in order to classify the unknown webpages in the next stage, i.e., the classification stage. The detailed steps of this stage are elaborated on as follows.

3.1.1 Features Extraction Module

In this module, each webpage will be analyzed to decide its critical attributes’ values by performing the following two steps: 1) preprocessing and 2) capturing attributes.

1) Preprocessing

In this step, each webpage should be converted into the HTML format. Then, spaces, punctuations, and stop-words of the words in webpage content will be removed. Moreover, we execute the Porter stemming algorithm\cite{28} to reduce repetitively arising words with different types. For example, “apple” and “apples” will be reduced to “appl” after stemming performed.

2) Capturing attributes

This step is designed to decide the values of critical attributes for each webpage. In order to filter pornographic webpages accurately, we will perform some judgments based on the suspicious elements of HTML structures, as well as the frequency of pornographic keywords. Based on the research of Lee et al.\cite{9} about the parts of the HTML code mostly likely to be dominated by pornographic keywords, we study and outline a number of critical attributes as shown in Table 2. For the sake of convenience, we use “XXX” to indicate the strings that pornographic keywords might appear. As described in Table 2, the critical attributes from No. 1 to No. 14 will be valued as 0 or 1 by checking whether the corresponding HTML elements of each webpage meet the judgment conditions. Moreover, these critical attributes will be examined whether they contain pornographic keywords via the keyword database. In this research, the pornographic keywords stored in the keyword database are collected from the website SafeSquid (http://www.safesquid.com/).

The 15th critical attribute (No. 15) is to check whether the webpage contains more than three keywords that can be found in the sliding window. The content of the sliding window will be introduced in the next module (concept drift scoring module). Moreover, the 16th critical attribute (No. 16) will also be handled in the concept drift scoring module.

3.1.2 Concept Drift Scoring Module

In this module, we use a window-based technique to estimate for the concept drift condition of pornographic keywords in each webpage. We propose a weighted “sliding window” to keep the keywords found recently in pornographic webpages. The size of the sliding window is defined to be 30 days. That is to say, the sliding window will store the keywords found frequently in the past 30 days. We assume that all of the webpages are entered into the proposed filtering system in the increasing order of the webpage’s date. For each webpage, this module will execute the following three tasks sequentially: 1) deciding the value of the 15th critical attribute; 2) calculating the score of concept drift; 3) updating the sliding window.

1) Deciding the value of the 15th critical attribute

Each incoming webpage will be checked whether it contains more than 3 keywords that have been recorded in the sliding window. The 15th critical attribute (as shown in Table 2) will be set as 1 if this incoming webpage contains more than 3 keywords that can be found in the sliding window, and set as 0 otherwise.
Table 2: Sixteen critical attributes used in this research

<table>
<thead>
<tr>
<th>No.</th>
<th>Critical attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Whether pornographic keywords exist in the HTML Tag of URL?</td>
<td>Set as 1 if the URL (URL://XXX) contains pornographic keywords, and set as 0 otherwise.</td>
</tr>
<tr>
<td>2</td>
<td>Whether pornographic keywords exist in the HTML Tag of title?</td>
<td>Set as 1 if the HTML Tag &lt;title&gt;XXX&lt;/title&gt; contains pornographic keywords should be set as 1, and set as 0 otherwise.</td>
</tr>
<tr>
<td>3</td>
<td>Whether pornographic keywords exist in the HTML Tag of link (I)?</td>
<td>Set as 1 if the HTML Tag &lt;a href=&quot;XXX&quot;&gt; contains pornographic keywords should be set as 1, and set as 0 otherwise.</td>
</tr>
<tr>
<td>4</td>
<td>Whether pornographic keywords exist in the HTML Tag of link (II)?</td>
<td>Set as 1 if the HTML Tag &lt;a href=&quot;XXX&quot;&gt; contains pornographic keywords should be set as 1, and set as 0 otherwise.</td>
</tr>
<tr>
<td>5</td>
<td>Whether pornographic keywords exist in the HTML Tag of metadata (I)?</td>
<td>Set as 1 if the HTML Tag &lt;meta name=&quot;author&quot; content=&quot;XXX&quot;&gt; contains pornographic keywords should be set as 1, and set as 0 otherwise.</td>
</tr>
<tr>
<td>6</td>
<td>Whether pornographic keywords exist in the HTML Tag of metadata (II)?</td>
<td>Set as 1 if the HTML Tag &lt;meta name=&quot;description&quot; content=&quot;XXX&quot;&gt; contains pornographic keywords should be set as 1, and set as 0 otherwise.</td>
</tr>
<tr>
<td>7</td>
<td>Whether pornographic keywords exist in the HTML Tag of metadata (III)?</td>
<td>Set as 1 if the HTML Tag &lt;meta name=&quot;keyword&quot; content=&quot;XXX&quot;&gt; contains pornographic keywords, and set as 0 otherwise.</td>
</tr>
<tr>
<td>8</td>
<td>Whether pornographic keywords exist in the HTML Tag of metadata (IV)?</td>
<td>Set as 1 if the HTML Tag &lt;meta name=&quot;description&quot; content=&quot;XXX&quot;&gt; contains pornographic keywords, and set as 0 otherwise.</td>
</tr>
<tr>
<td>9</td>
<td>Whether pornographic keywords exist in the HTML Tag of hyperlink (I)?</td>
<td>Set as 1 if the HTML Tag &lt;a href=&quot;XXX&quot;&gt; contains pornographic keywords, and set as 0 otherwise.</td>
</tr>
<tr>
<td>10</td>
<td>Whether pornographic keywords exist in the HTML Tag of hyperlink (II)?</td>
<td>Set as 1 if the HTML Tag &lt;a href=&quot;XXX&quot;&gt; contains pornographic keywords, and set as 0 otherwise.</td>
</tr>
<tr>
<td>11</td>
<td>Whether pornographic keywords exist in the HTML Tag of image (I)?</td>
<td>Set as 1 if the HTML Tag &lt;img src=&quot;XXX&quot;&gt; contains pornographic keywords, and set as 0 otherwise.</td>
</tr>
<tr>
<td>12</td>
<td>Whether pornographic keywords exist in the HTML Tag of image (II)?</td>
<td>Set as 1 if the HTML Tag &lt;img src=&quot;XXX&quot;&gt; contains pornographic keywords, and set as 0 otherwise.</td>
</tr>
<tr>
<td>13</td>
<td>Whether pornographic keywords exist in the HTML Tag of image (III)?</td>
<td>Set as 1 if the body (the textual content of the webpage) contains more than 5 pornographic keywords, and set as 0 otherwise.</td>
</tr>
<tr>
<td>14</td>
<td>Whether there exist more than 5 pornographic keywords in the texts defined by the HTML Tag of body?</td>
<td>Set as 1 if the webpage contains more than 3 keywords that can be found in the sliding window, and set as 0 otherwise.</td>
</tr>
<tr>
<td>15</td>
<td>Whether this webpage contains more than 3 keywords of the sliding window?</td>
<td>Computed by the concept drift scoring module to estimate for this webpage's concept drift condition of pornographic keywords.</td>
</tr>
<tr>
<td>16</td>
<td>The concept drift score of this webpage</td>
<td></td>
</tr>
</tbody>
</table>

2) Calculating the score of concept drift

Suppose that a new webpage, denoted as \( w \), is entered into the system. Moreover, we assume that this webpage includes \( k \) keywords which can be found in the sliding window. Without loss of generality, we let these keywords be denoted as \( \text{Keyword}(i) \) for \( 1 \leq i \leq k \) where \( k \) is a positive integer. The following functions are necessary to calculate the concept drift score of this webpage \( w \):

a) \( \text{Keyword Frequency}(i) \): Keyword’s frequency is a significant criterion of concept drift, hence we propose this function to accumulate the found times of \( \text{Keyword}(i) \) since it was stored into the sliding window. Whenever the same keyword, \( \text{Keyword}(i) \), is found in the incoming webpage, the value of \( \text{Keyword Frequency}(i) \) will be increased by 1.

b) \( \text{Webpage Date}(w) \): This function records date time of the current webpage \( w \).

c) \( \text{Access Date}(i) \): This function records the latest date time that \( \text{Keyword}(i) \) was found in an incoming webpage previous to \( w \).

d) \( \text{Time Differ}(i) \): This function is to calculate the sustained time owing to the latest date time that \( \text{Keyword}(i) \) appeared in an incoming webpage previous to \( w \). It is computed by using the following formula:

\[
\text{Time Differ}(i) = \text{Webpage Date}(w) - \text{Access Date}(i)
\]

e) \( \text{Time Weight}(i) \): The concentrated appearance of a new keyword usually implies occurring of concept drift. The purpose of this function is to compute the weighted value of estimating the degree of concentrated appearance for each keyword. Obviously, the concentrated appearance of a keyword will result in the smaller value of \( \text{Time Differ}(i) \), therefore, its weighted value should be greater. Assume that the units of measurement for \( \text{Time Differ}(i) \) are days, the weighted value will be calculated by

\[
\text{Time Weight}(i) = 30 - \text{Time Differ}(i)
\]

Now, we can compute \( \text{Concept Drift Score}(w) \), the score of concept drift of the current webpage \( w \), by using the following equation:

\[
\text{Concept Drift Score}(w) = \sum_{i=1}^{k} (\text{Keyword Frequency}(i) \times \text{Time Weight}(i)).
\]

Thus, the value of the 16th critical attribute (No. 16 of Table 2) of each incoming webpage can be acquired.

3) Updating the sliding window

Whenever a pornographic webpage is entered into the system, each keyword found in this webpage will be considered as a candidate for storing into the sliding window. In our method, there are three cases of updating the content of sliding window: (I) adding new keywords during the training stage; (II) updating the values of related functions; (III) adding new keywords by the incremental learning module of the classification stage.
Case (I) occurs in the training stage. If a pornographic training webpage is entered into the system, we will check each keyword found in this webpage whether it has been stored in the sliding window. If the keyword is not recorded in the sliding window, we store it into the sliding window and initialize the values of its related functions as mentioned above.

Case (II) occurs whenever a keyword that has been stored in the sliding window is found in an incoming webpage. If a keyword (say, Keyword(i)) stored in the sliding window is found in an incoming webpage, we will modify its related functions’ values as follows: 1) adding 1 to its frequency value (i.e., Keyword_Frequency(i)); 2) changing its latest accessed time (i.e., Access_Date(i)) into the date time of this incoming webpage.

Case (III) occurs in the classification stage. If an unknown webpage is classified as pornographic, the incremental learning module of the classification stage will check each keyword found in this webpage; if this keyword has not been recorded in the sliding window yet, it will be stored into the sliding window and the values of its related functions will be initialized.

Therefore, the webpage with recurring keywords will gain greater weights in calculating its concept drift score (as mentioned earlier). Thus, this module will help improve the accuracy of identifying pornographic webpages by adding recurring keywords into the sliding window. Note that the sliding window store the keywords found frequently within the past 30 days in our method. Each keyword with the Time_Differ(i) more than 30 days will be deleted automatically.

3.1.3 Decision Tree Data Mining Module

Now, the values of all critical attributes of each training webpage are acquired. In this module, the decision tree data mining algorithm C4.5 will be applied to all training webpages. The C4.5 algorithm will construct a decision tree and bring out the potential association rules of “if-then” pattern between 16 critical attributes (as described in Table 2) and target attribute (webpage’s type: Pornographic or legitimate). Then, this module will score each association rule by using formulas based on the values of the degree of support and the degree of purity, which have been defined in Section 2.

Given any association rule $R$, we assume that the associated leaf node is $C$, and $p$ is the number of webpages whose target attribute’s value is “$P$” (pornographic) in node $C$. To describe the scoring formula for the rule, the following four important functions are necessary.

1) $P\_Degree(R)$: This function records the “pornographic degree” for rule $R$, which implies the “intensity” of rule $R$ to classify an unknown webpage as “$P$” (pornographic). Let $|C|$ be the number of webpages contained in node $C$, this function is defined as follows:

$$P\_Degree(R) = \frac{Purity(C)}{} \text{ if Label}(C) = "P";$$

$$P\_Degree(R) = \frac{p/|C| \times 100\%}{\text{ otherwise}}.$$  

2) Rule_Support($R$): This function records the degree of support for rule $R$, which is defined as follows:

$$\text{Rule\_Support}(R) = \text{Support}(C).$$

3) $W$(Rule_Support($R$)): This function records the weighted value for Rule_Support($R$). We compute the values of the degree of support for all rules and assume that Rule_Support$_{\text{MAX}}$ is the maximum one and Rule_Support$_{\text{MIN}}$ is the minimum one. Then this function can be defined as follows:

$$W(\text{Rule\_Support}(R)) = \frac{\text{Rule\_Support}(R)}{\text{Rule\_Support}_{\text{MAX}} + \text{Rule\_Support}_{\text{MIN}}} \times 100\%.$$  

4) Score(Rule_Support($R$)): The function calculates score for Rule_Support($R$). We compute the weighted values of rule support for all rules by using above formula of $W(\text{Rule\_Support}(R))$, and assume that Weight$_{\text{MAX}}$ is the maximum one and Weight$_{\text{MIN}}$ is the minimum one. Then, this function can be calculated as follows:

$$\text{Score}(\text{Rule\_Support}(R)) = \frac{W(\text{Rule\_Support}(R)) - \text{Weight}_{\text{MIN}}}{\text{Weight}_{\text{MAX}} - \text{Weight}_{\text{MIN}}} \times 100\%.$$  

Now, we can describe the formula of computing score for rule $R$. Let the score of rule $R$ be Score($R$), which is defined as follows:

$$\text{Score}(R) = (70\% \times P\_Degree(R) + 30\% \times \text{Score}(\text{Rule\_Support}(R))) \times 100\%.$$  

Then, we compute scores for all association rules. We choose the minimum score from the rules whose values of pornographic degree are more than 80% (i.e., $P\_Degree(R) \geq 80\%$), and set it as the threshold, denoted as “t_value”, for judging whether an unknown webpage is pornographic. The results obtained by above steps will be stored in the rule database and accessed by the following stage (classification stage) to classify unknown webpages accordingly.

3.2 Classification Stage

With the rule database computed by the training stage, now the classification stage will classify unknown webpages to be either pornographic or legitimate.

The first step of this stage is to examine unknown webpages; the features extraction module will compute values of critical attributes from No. 1 to No. 14 (as shown in Table 2) and the concept drift scoring module will compute the 15th and 16th critical attributes for each unknown webpage. Then, the webpage scoring module will use these 16 critical attributes to inquire the rule database and in order to find the score for each unknown webpage. Therefore, we can judge whether this unknown webpage is pornographic by comparing its score with the threshold of classification (i.e., t_value). Finally, if an unknown webpage is classified as pornographic, the incremental learning module will be started up to learn new pornographic keywords into the sliding windows and keyword
database. The detailed process of this stage can be partitioned into the modules as follows.

3.2.1 Features Extraction Module

This task of this features extraction module is identical to that of the training stage. In this module, each unknown webpage will be analyzed to extract its critical attributes’ values by performing the following steps: 1) preprocessing; 2) capturing attributes.

Like the steps of the features extraction module in the training module, the step “preprocessing” is to convert each unknown webpage into HTML format and reduce its noise in webpage content. Then, the step “capturing attributes” will compute the values of 14 critical attributes (from No. 1 to No. 14 as shown in Table 2) for each unknown webpage.

3.2.2 Concept Drift Scoring Module

For each unknown webpage, this module will execute the following three tasks: 1) deciding the value of the 15th critical attribute; 2) calculating the score of concept drift; 3) updating the sliding window.

1) Deciding the value of the 15th critical attribute

The 15th critical attribute (as shown in Table 2) will be set as 1 if this unknown webpage contains more than 3 keywords that can be found in the sliding window, and set as 0 otherwise.

2) Calculating the score of concept drift

The process of this task is the same as that of the concept drift scoring module in the training stage, which will compute the score of concept drift for each unknown webpage to estimate for the condition of concept drift in this webpage. Thus, the 16th critical attribute (No. 16 of Table 2) of this unknown webpage will be extracted.

3) Updating the sliding window

Whenever any keyword of sliding window is found in an unknown webpage, we will modify its related functions’ values. Let this found keyword be Keyword(i). We execute the following two modifications for Keyword(i): 1) adding 1 to its frequency value, i.e., Keyword_Frequency(i) and 2) rewriting its latest accessed time, Access_Date(i), as the date time of this unknown webpage.

3.2.3 Webpage Scoring Module

Now, all the 16 critical attributes (as described in Table 2) of this unknown webpage have been extracted. According to these critical attributes’ values, this unknown webpage will dovetail with some association rule stored in the rule database built. Assume that this dovetailed rule is R. Then, we designate the score of this unknown webpage as Score(R), which is defined as the formula described in the decision tree data mining module of the training stage. This unknown webpage will be classified as pornographic if its score is not lower than the classification threshold (i.e., Score(R) ≥ t_value), and classified as legitimate otherwise.

3.2.4 Incremental Learning Module

At this time, each unknown webpage has been classified to be either pornographic or legitimate. If an unknown webpage is classified as pornographic, this module will check each keyword found in this webpage by using the supervised learning technique.

If any keyword found in this pornographic webpage has not been recorded in the sliding window yet, it will be stored in the sliding window and the values of its related functions will be initialized. Moreover, we will check whether this pornographic webpage contains any new pornographic keyword that is not recorded in the keyword database. If so, each of new pornographic keywords will be stored in the keyword database.

Therefore, this module will help improve the accuracy of identifying pornographic webpages by learning incrementally new pornographic keywords to enrich the keyword database and adding recurring keywords into the sliding window to ameliorate the detection of concept drift. Hence, our filtering method can better satisfy the current time concept when reviewing and recording the webpage content characteristics to allow the filtering system better fit for the dynamic environment.

4. Experimental Results

4.1 Experimental Design

In this study, we used the decision tree data mining technique to analyze pornographic webpage content information to speed up the filtering velocity of pornographic webpage filter. The mechanisms to judge the occurrence of concept drift and carry out the incremental learning were also included to improve webpage filter precision. Then, we designed some experiments to confirm the accuracy and efficiency of the proposed filtering method.

The webpage data used in the experiments was from the pornographic webpage database provided by the website of URLBlack.com (http://urlblacklist.com/). We collected 600 pornographic webpage files each month during the period from February to May 2011 and a total of 2400 pornographic webpage files were collected. Let these webpages be sorted in the increasing order of the date time. According to the date time, we classified these webpages into four groups such that “Month (I)” contained the webpages of February, “Month (II)” contained the webpages of March, “Month (III)” contained the webpages of April, and “Month (IV)” contained the webpages of May. The purpose of this classification was to test whether filtering precision of each month could be improved due to the incremental learning mechanisms of the keyword database and sliding window. For the sake of experimental fairness, the monthly pornographic webpage and normal webpage file data were mixed in the proportion of 1:1. In the training stage, we selected randomly 300 pornographic webpages of each month as the training data, which were mixed with 300 legitimate webpages. In the classification stage, we took the remaining 300 pornographic webpages of
each month as the unknown data, which were mixed with 300 legitimate webpages. Moreover, these webpages of each month would be applied to the experiments in the increasing order of date time.

To measure the experimental performance of the proposed systematic method in terms of webpage filtering accuracy, in addition to the indexes: “Precision” and “Recall”, which were commonly used for document classification, we calculated the harmonic mean “F-measure”. As shown in Table 3, we used the decision confusion matrix to explain the calculation equations listed below. The four cases $A$, $B$, $C$, and $D$ in Table 3 are all recorded by the number of webpages.

<table>
<thead>
<tr>
<th>Table 3: Four cases of judgment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term</td>
</tr>
<tr>
<td>To be judged as pornographic</td>
</tr>
<tr>
<td>To be judged as legitimate</td>
</tr>
</tbody>
</table>

1) Accuracy rate
It refers to the ratio of webpages classified correctly. It is the probability of legitimate webpages being classified as legitimate and pornographic webpages being filtered as pornographic. The equation is as shown below:

$$\text{Accuracy} = \frac{A+D}{A+B+C+D}.$$  

2) Precision rate
It refers to the ratio of the webpages identified correctly in the webpages classified as the certain category, representing filter’s capability to classify correctly such category of webpages. The “PornographicPrecision” is computed from the viewpoint of identifying the pornographic webpages; and “LegitimatePrecision” is computed from the viewpoint of identifying the legitimate webpages. Finally, “Precision” is set as the average of the addition of PornographicPrecision and LegitimatePrecision, and the equations are as follows:

$$\text{PornographicPrecision} = \frac{A}{A+B}$$
$$\text{LegitimatePrecision} = \frac{D}{C+D}$$
$$\text{Precision} = \frac{(\text{PornographicPrecision} + \text{LegitimatePrecision})}{2}.$$  

3) Recall rate
It refers to the probability of identifying correctly the certain category of webpages. The “PornographicRecall” is defined as the probability of judging pornographic webpages as pornographic ones. And the “LegitimateRecall” is defined as the probability judging correctly legitimate webpages. Finally, “Recall” is set as the mean of PornographicRecall and LegitimateRecall, and the equations are as follows:

$$\text{PornographicRecall} = \frac{A}{A+C}$$
$$\text{LegitimateRecall} = \frac{D}{B+D}$$
$$\text{Recall} = \frac{(\text{PornographicRecall} + \text{LegitimateRecall})}{2}.$$  

4) F-measure
Regarding the F-measure of a harmonic mean of Precision and Recall, the equation is as follows:

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$  

To confirm the efficiency of our filtering method proposed in this paper, we performed the following three experiments:

- **Experiment (A):** This experiment performed the proposed method of this paper without using the incremental learning module. In other words, the new keywords of pornographic webpages identified in the classification stage would not be stored into either the keyword database nor the sliding window.
- **Experiment (B):** This experiment performed the proposed method with using the incremental learning module to increase the keyword database only. In other words, we stored the new pornographic keywords of pornographic webpages identified in the classification stage into the keyword database only.
- **Experiment (C):** This experiment performed the proposed method with using the incremental learning module to increase both the keyword database and sliding window. Namely, for each pornographic webpage identified by the classification stage, we not only stored its newly found keywords into the sliding window but also stored its newly found pornographic keywords into the keyword database. Therefore, the proposed method in this paper would better satisfy the current time concept to allow the filtering system better fit for the dynamic environment.

### 4.2 Experimental Results

Now we discuss the results of the three experiments. As shown in Table 4, the averaged accuracy rate of Experiment (A) (performed the proposed method without using the incremental learning module) was 91.80%, which was the lowest among the three experiments. Without incremental learning, the content of keyword database was stationary, i.e., the number of pornographic keywords stored in the keyword database was fixed. However, the webpages’ contents in the real world were not fixed. Hence, the filter’s capabilities of filtering pornographic webpages were quite limited. Many pornographic webpages of the unknown ones could not be accurately judged and thus the accuracy rate was relatively low.

<table>
<thead>
<tr>
<th>Table 4: Accuracy rate of the three experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
</tr>
<tr>
<td>Month (I)</td>
</tr>
<tr>
<td>Month (II)</td>
</tr>
<tr>
<td>Month (III)</td>
</tr>
<tr>
<td>Month (IV)</td>
</tr>
<tr>
<td>Average</td>
</tr>
</tbody>
</table>

On the other hand, the proposed method could add new pornographic keywords increasingly into the keyword database.
by using the incremental learning module. Thus, the averaged accuracy rate of Experiment (B) was raised to 94.85%, suggesting the incremental learning technique could significantly improve the accuracy of filtering pornographic webpages.

Moreover, Experiment (C) used the incremental learning module to detect recurring keywords and store them into the sliding window to help evaluate the concept drift condition of webpage, hence, its accuracy rate was the highest among three experiments at 97.06%.

The experiments results showed that the accuracy rate of our filtering method proposed in this paper would improve over a number of months. Note that the accuracy rate was the ratio of webpages being accurately filtered, namely, legitimate webpages were classified by the proposed method as legitimate ones while pornographic webpages were classified by the proposed method as pornographic ones. Without any incremental learning, as the Experiment (A) used the fixed keyword database to classify the unknown webpages, the resulting accuracy rate was obviously lower. The highest accuracy rate in the case of having no incremental learning was 93.33%. By using the incremental learning technique to add new pornographic keywords into the keyword database, the accuracy rate of Experiment (B) significantly increased over the four months, suggesting that incremental learning can substantially contribute to the identification of pornographic webpages. The Experiment (C) used incremental learning technique coupled with the sliding window to store the recurring keywords, which would help evaluate the concept drift condition of webpage. As the experimental results suggested, the accuracy rate of the Experiment (C) could be improved outstandingly and the highest value reached 98.33%.

Changes of the precision rate by month were shown in Table 5. Note that the precision rate was the percentage of webpages identified correctly in the webpages judged as the certain category. Without any incremental learning, the precision rate of Experiment (A) was obviously lower than those of Experiments (B) and (C). With incremental learning, by unceasingly adding new pornographic keywords into the keyword database, the precision rate of Experiment (B) significantly improved, and the highest value was 96.77%. In Experiment (C), the weighted sliding window was used to record the recurring keywords. Then, the precision rate of Experiment (C) reached 98.67% in Month (IV) and was the highest among these three experiments. Hence, compared with Experiment (B), the proposed method with both incremental learning and detection of concept drift would further enhance the precision obviously.

The monthly changes in the recall rate of the proposed method were shown in Table 6. Note that the recall rate was the ratio of identifying correctly the certain category of webpages. Without any incremental learning, the recall rate of Experiment (A) was the lowest, and the highest value was 92.76%; with the help of the new pornographic keywords added to the keyword database, the recall rate of Experiment (B) was significantly improved due to the technique of incremental learning, and the highest value was 97.99%; in the case of integrating the sliding window with incremental learning, the recall rate of Experiment (C) was slightly improved and the highest value was 98.01%.

Table 7 illustrated the monthly changes in F-measure, the harmonic mean of the precision rate and recall rate, of identifying unknown webpages by the proposed method. Without any incremental learning, the F-measure of Experiment (A) was the lowest. With incremental learning of adding new pornographic keywords into the keyword database to help identify the pornographic webpages, the F-measure of Experiment (B) significantly improved and reached 97.38% in Month (III). It could be observed that the incremental learning technique would lead to significant difference in identifying unknown webpages. Moreover, in Experiment (C), the incremental learning technique was used in combination with the weighted sliding window to store the recurring keywords, which would help improve the effectiveness of identifying pornographic webpages by giving greater weights to the webpages with recurring keywords. Therefore, the F-measure of Experiment (C) was the highest among these three experiments, and it reached the highest value of 98.34% in Month (IV). Hence, incremental learning concept drift can effectively help the classification of unknown webpages.

5. Conclusion

In this paper, we used the decision tree data mining technique to track the concept drift of pornographic webpages. Pornographic webpages can easily change with the real world to produce new information to generate concept drift. To allow the filtering method proposed in this paper to suit the dynamic
environment and improve its accuracy and performance, we proposed the weighted sliding window to store the recurring keywords, which would help evaluate the concept drift condition of unknown webpages. And we designed an incremental learning mechanism to learn newly found pornographic keywords, which would substantially contribute to the identification of pornographic webpages. According to the experimental results, our method coupled with the incremental learning technique and detection of concept drift had the accuracy rate up to 98.33% in the actual classification of pornographic webpages.

The proposed webpage filter in this paper can only filter English webpages. Due to the development of the Internet, pornographic webpage information is increasingly internationalized and diversified. Therefore, we suggest that the future development direction is the development of filters that can identify pornographic webpages of multiple languages such as Chinese, Japanese, and Korean.

References


Jyh-Jian Sheu is currently an associate professor with the College of Communication, National Chengchi University, Taipei. He received his B.B.A. degree in management information systems from National Chengchi University, and his M.S. and Ph.D. degrees in computer and information science from National Chiao Tung University, Hsinchu. His primary research interests include data mining, Internet security, and big data.