# 科技部補助專題研究計畫成果報告
# 期末報告

## 一個能兼具相似度與差異度計算以及再學習機制的有效率電子文件辨識方法：以色情及醫學網頁辨識為例

計 畫 類 別 ： 個別型計畫
計 畫 編 號 ： MOST 103-2410-H-004-112-
執 行 期 間 ： 103年08月01日至104年07月31日
執 行 單 位 ： 國立政治大學傳播學院

計 畫 主 持 人 ： 許志堅

計畫參與人員： 碩士班研究生-兼任助理人員：宋伯謙
大專生-兼任助理人員：鄭御廷

處 理 方 式 ：
1.公開資訊：本計畫涉及專利或其他智慧財產權，2年後可公開查詢
2.「本研究」是否已有嚴重損及公共利益之發現：否
3.「本報告」是否建議提供政府單位施政參考：否

中 華 民 國 104 年 10 月 30 日

中 文 摘 要 ： 本研究提出一個系統化而且有效的的電子文件辨識方法，除了計算相似性，也能分析其差異性以避免誤判。我們以色情網頁與醫學相關網頁為例，利用機器學習方法當中的決策樹資料探勘演算法來進行不同類型網頁所擁有的特徵屬性的知識學習，尋求其關聯式規則作為未知文件之判斷。並且具備以下特色：

一、色情資訊與醫學資訊的個別知識與混合知識的關聯式規則分析：我們分析色情資訊以及醫學相關資訊可供比對與過濾的特徵，分別設計三種不同類型資料進行決策樹計算：(1)色情網頁決策樹分析：針對色情網頁的特徵進行訓練與計算，尋求單獨過濾色情網頁時的關聯式規則；(2)醫學網頁決策樹分析：針對醫學網頁的特徵進行訓練與計算，尋求單獨辨識醫學網頁時的關聯式規則；(3)色情網頁與醫學網頁混合資料決策樹分析：針對混和色情、醫學資訊之網頁特徵進行訓練與計算，找出在二者資訊可能同時並存的情形下，如何辨識雙方的關聯式規則。除了希望提高對於色情網頁的過濾能力之外，也能正確辨識醫學相關網頁，避免產生誤判與混淆。

二、再學習機制：色情網頁內容可能隨著時間或是當時熱門的事件而有變化、不斷推陳出新，而造成過濾上困難。我們利用機器學習的方式設計一套“再學習”機制，以獲取色情文件特徵值的動態關鍵字變化。

三、提出兼具效率與正確性的色情網頁與醫學(含性教育)網頁過濾機制：本研究以特徵值擷取為基礎，避免對圖片進行費時的分析、同時避免耗時的語意分析計算；運用ID3演算法來建構一個系統化的具備效率的過濾機制，並且獲得較高的過濾準確性。

中 文 關 鍵 詞 ： 文件分類、資料探勘、決策樹、色情網頁過濾

英 文 摘 要 ： In this study, we apply decision tree data mining technique to basic attributes of porn sites to analyze the association rules for indentifying an unknown web site to be either legitimate or porny. We focus on web’s context and apply decision tree data mining technique to analyze the association rules for medical pages and pornographic pages. Then we propose a systematic method to accurately identify an unknown web to be either porny or legitimate. There are three major parts in this project, which are described as follows:

(I)To compute associative rules for medical page and pornographic page respectively:
We design three kinds of rule database for the computation of decision tree：(1) Database of pornographic page; (2) Database of medical page; (3) Database of mix of medical page and pornographic page.

(II)The re-learning mechanism:

Since the keywords of pornographic page are constantly changing, we construct a re-learning mechanism of recording new pornographic keywords.
(III) An effective filtering mechanism with outstanding ability of recognizing porn sites:
Without handling pictures and semantic analysis, we propose our effective filtering method by applying associative rules and keywords only.

英文關鍵詞：Filtering Porn Sites, Decision Tree, Data Mining, Document Classification

# 行政院科技部補助專題研究計畫 ■成果報告 □期中進度報告

# 一個能兼具相似度與差異度計算以及再學習機制的有效率電子文件辨識方法：以色情及醫學網頁辨識為例

計畫類別：■個別型計畫　□整合型計畫
計畫編號：MOST 103-2410-H-004 -112
執行期間：103 年 8 月 1 日至 104 年 7 月 31 日

執行機構及系所：國立政治大學傳播學院

計畫主持人：許志堅
共同主持人： 朱克聰
計畫參與人員：

成果報告類型(依經費核定清單規定繳交)：■精簡報告　□完整報告

本計畫除繳交成果報告外，另須繳交以下出國心得報告：
□赴國外出差或研習心得報告
□赴大陸地區出差或研習心得報告
□出席國際學術會議心得報告
□國際合作研究計畫國外研究報告

處理方式：除列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年■二年後可公開查詢

中　華　民　國　104　年　7　月　31　日

# 一、摘要

**中文摘要：**

本研究提出一個系統化而且有效的的電子文件辨識方法，除了計算相似性，也能分析其差異性以避免誤判。我們以色情網頁與醫學相關網頁為例，利用機器學習方法當中的決策樹資料探勘演算法來進行不同類型網頁所擁有的特徵屬性的知識學習，尋求其關聯式規則作為未知文件之判斷。並且具備以下特色：

- 色情資訊與醫學資訊的個別知識與混合知識的關聯式規則分析：我們分析色情資訊以及醫學相關資訊可供比對與過濾的特徵，分別設計三種不同類型資料進行決策樹計算：(1)色情網頁決策樹分析：針對色情網頁的特徵進行訓練與計算，尋求單獨過濾色情網頁時的關聯式規則；(2)醫學網頁決策樹分析：針對醫學網頁的特徵進行訓練與計算，尋求單獨辨識醫學網頁時的關聯式規則；(3)色情網頁與醫學網頁混合資料決策樹分析：針對混和色情、醫學資訊之網頁特徵進行訓練與計算，找出在二者資訊可能同時並存的情形下，如何辨識雙方的關聯式規則。除了希望提高對於色情網頁的過濾能力之外，也能正確辨識醫學相關網頁，避免產生誤判與混淆。

- 再學習機制：色情網頁內容可能隨著時間或是當時熱門的事件而有變化、不斷推陳出新，而造成過濾上困難。我們利用機器學習的方式設計一套“再學習”機制，以獲取色情文件特徵值的動態關鍵字變化。

- 提出兼具效率與正確性的色情網頁與醫學(含性教育)網頁過濾機制：本研究以特徵值擷取為基礎，避免對圖片進行費時的分析、同時避免耗時的語意分析計算；運用 ID3 演算法來建構一個系統化的具備效率的過濾機制，並且獲得較高的過濾準確性。

**關鍵詞：**文件分類、資料探勘、決策樹、色情網頁過濾

**英文摘要：**

In this study, we apply decision tree data mining technique to basic attributes of porn sites to analyze the association rules for indentifying an unknown web site to be either legitimate or porny. We focus on web's context and apply decision tree data mining technique to analyze the association rules for medical pages and pornographic pages. Then we propose a systematic method to accurately identify an unknown web to be either porny or legitimate.

There are three major parts in this project, which are described as follows:

I. **To compute associative rules for medical page and pornographic page respectively.**
We design three kinds of rule database for the computation of decision tree：(1) Database of pornographic page; (2) Database of medical page; (3) Database of mix of medical page and pornographic page.

II. **The re-learning mechanism.**
Since the keywords of pornographic page are constantly changing, we construct a re-learning mechanism of recording new pornographic keywords.

III. **An effective filtering mechanism with outstanding ability of recognizing porn sites.**
Without handling pictures and semantic analysis, we propose our effective filtering method by applying associative rules and keywords only.

**Keywords:** Filtering Porn Sites, Decision Tree, Data Mining, Document Classification

# 二、報告內容

本研究成果已經撰寫為英文，正準備投稿到 SCI 國際期刊，以下節錄本研究的第四章研究架構與第五章實驗成果、第六章結論與建議之部分：

## 4. An efficient pornographic websites filtering mechanism

The objective of this research is to filter pornographic web pages, namely, classifying the web pages into the two target categories of pornographic and non-pornographic web pages. 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 general (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 or 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 2, 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.

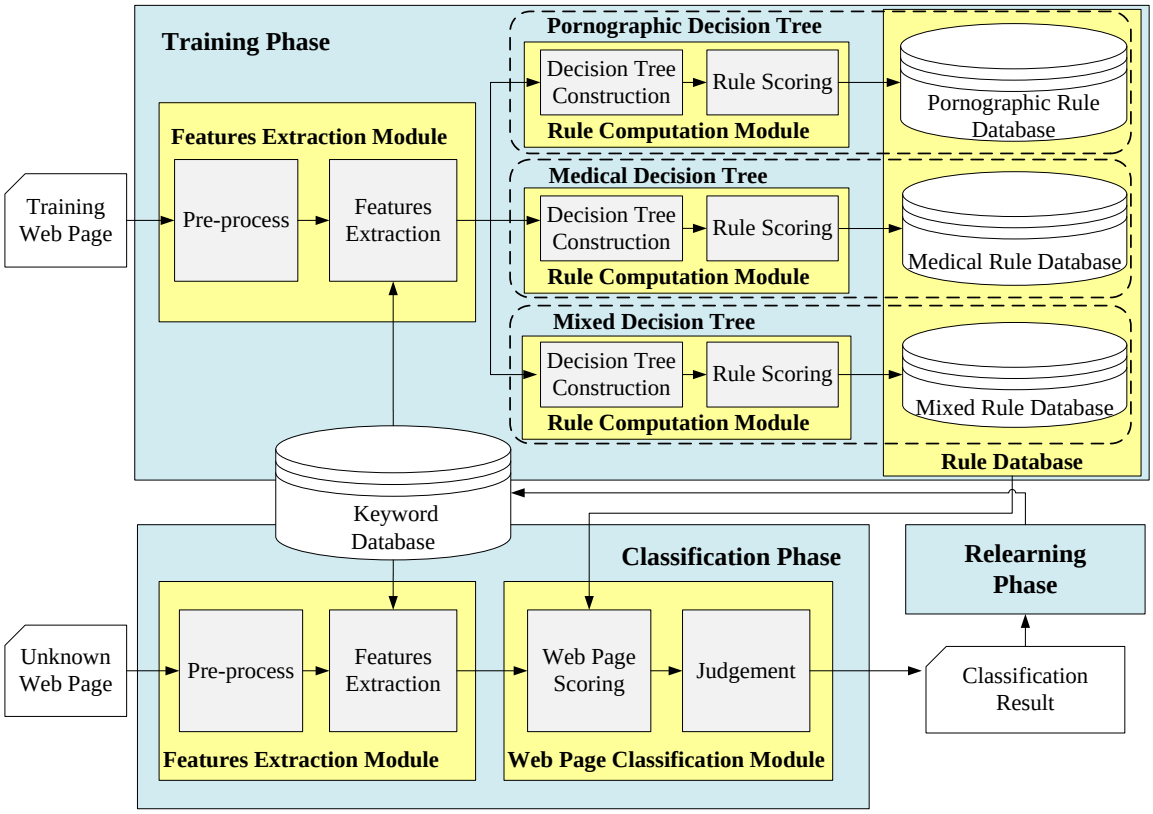


Figure 2. Structure of the proposed method

## 4.1 The Training Phase

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

Each training web page must be examined by the Features Extraction Module to extract its critical features first in this phase. Then, training web pages will be transmitted to Rule Computation Modules in order to construct decision tree and calculate the association rules. As shown in Figure 2, we construct three decision trees (Pornographic Decision Tree, Medical Decision Tree, and Mixed Decision Tree) and apply three copies of Rule Computation Module individually to acquire various association rules, which will be stored respectively into the corresponding rule databases (Pornographic Rule Database, Medical Rule Database, and Mixed Rule Database).

In the Pornographic Decision Tree, the pornographic training data (web pages) and normal training data (web pages) will be used as input data for the Rule Computation Module, which will compute the association

rules of distinguishing pornographic websites from normal websites. Then, the medical training data and normal training data will be used as input data for the Rule Computation Module in the Medical Decision Tree to compute the association rules of distinguishing medical websites from normal websites. Moreover, the mix of medical training data and pornographic training data will be used as input for the Rule Computation Module in the Mixed Decision Tree to compute the association rules of distinguishing between medical websites and pornographic websites.

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

### 4.1.1 Features Extraction Module

In the Features Extraction Module, each web page will be analyzed and its critical characteristics of each web page will be extracted by applying the following two steps: (1) Pre-process; (2) Features extraction.

The first step is pre-process. Each web page should first be converted into the HTML format. It is common for pornographic websites to contain indecent or erotic keywords, while medical websites are generally featured with medical terms or disease-related keywords. In order to catch these keywords, the web pages should be converted into HTML format, which will be examined by the second step such that pornographic or medical keywords could be found in the HTML structures.

The second step is features extraction. In this step, the critical features concerning the identification of medical and pornographic web pages will be extracted. This step is designed to discern the web pages features, 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. Several check elements are outlined in Table 2, which will later serve as the Critical Attributes for the computation of association rules.

All these Critical Attributes are valued by 0, 1, and 2, it should be checked whether the corresponding HTML elements meet the setting conditions according to Table 2. Note that the Critical Attributes will be examined whether they contain pornographic and medical keywords via the Keyword Database. In this research, the pornographic keywords used are those collected from the website SafeSquid [31], and the medical keywords used in this research are those collected from the website MedlinePlus [26]. 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.

Moreover, 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 should be used to build the decision tree in the Decision Tree Construction Module.

Table 2. The Critical Attributes used in this study

| Type | No. | Description of the Critical Attributes | Judgment condition |
|---|---|---|---|

| | | | |
|---|---|---|---|
| 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; |

### 4.1.2 Rule Computation Module

As shown in Figure 2, 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 input data for building the Pornographic Decision Tree are those critical characteristics of the pornographic training pages and normal training pages that have been picked out in the previous module. In building the Medical Decision Tree, the input data are the critical characteristics of the medical training pages and normal training pages extracted in the previous module. Moreover, the input data for constructing the Mixed Decision Tree are the extracted critical characteristics of the mix of pornographic, medical, and training pages in the previous module.

That is, 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 *C* is the corresponding leaf node, 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" and $n_M$ is the number of data instances concerning the Target Attribute's value is "M" in *C*. The following three important 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:

$$PornDegree(R) = Purity(C) \quad if \quad Label(C) = "P";$$

and $$PornDegree(R) = (\frac{n_P}{|C|}) * 100\% \quad otherwise .$$

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

$$MedicalDegree(R) = Purity(C) \ \ if \ \ Label(C) = "M";$$

$$\text{and} \ \ MedicalDegree(R) = (\frac{n_M}{|C|}) * 100\% \ \ otherwise.$$

Finally, we introduce the formulas of computing pornographic score and medical score for rule *R* respectively: $PornScore(R)$ and $MedicalScore(R)$. 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:

$$PornScore(R) = (1 \times PornDegree(R) + 0.3 \times Weight(R)) \times 100;$$

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

By applying the formulas mentioned above, pornographic score and medical score of all rules can be acquired. Then, all rules of three decision trees are 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 according pornographic scores and medical scores computed above.

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 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.

## 4.2 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 2, 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.

### 4.2.1 Features Extraction Module

The task of Features Extraction Module is the same as that of Training Phase. Each unknown web page will be processed by the following two steps: (1) Pre-process; (2) Features extraction. The first step is to perform pre-process for each unknown web page. In this step, each unknown web page will be converted into the HTML format. Then, the second step is to extract the critical features concerning the identification of medical and pornographic web pages by examining the HTML structure of each unknown web page. By checking the elements outlined in Table 2, 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 classify this web page as pornographic, medical, or normal.

### 4.2.2 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 judge the category of each unknown web page.

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 its 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.

## 4.3 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 in the Keyword Database.

# 5. Experimental design and results

The objective of this research is to distinguish pornographic web pages from non-pornographic ones and avoid misjudging medical web pages as pornographic. 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 were classified into two categories: 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 3 to estimate the classification results. The proposed method in this research was designed to classify pornographic web pages correctly. As shown in Table 3, TP means that pornographic web pages are classified correctly as pornographic; TN means that normal web pages are classified as normal web pages. FN and FP refers to misjudgments. To be specific, FN means non-pornographic web pages are misjudged as pornographic and FP means pornographic web pages are misjudged as non-pornographic. Note that the non-pornographic web pages here include normal web pages and medical web pages.

Table 3. Four cases of judgement

| Classification / In reality | **Pornographic web pages** | **Non-pornographic web pages** |
|---|---|---|
| **Pornographic web pages** | TP (true positive) | FP (false positive)<br>Type Ⅰ error |
| **Non-pornographic web pages** | FN (false negative)<br>Type Ⅱ error | TN (true negative) |

The four values of TP, FP, FN, TN will be selected to constitute the three efficacy assessment indexes of “Accuracy”, “Precision” and “Recall”, which are used to evaluate the filtering accuracy concerning pornographic web pages. 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. The higher the result of this index, the better the judgment effects of the filtering mechanism. This index is calculated through the following formula:

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

Precision is used to evaluate the precision of the classification results concerning pornographic web pages, namely the proportion of pornographic web pages among all the web pages that are judged to be pornographic in nature. The higher this index, the better will be the classification effect and the lower the rate of misjudging non-pornographic web pages as pornographic ones. It is calculated through the following

formula:

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

Recall is used to evaluate the recall rate of the forecasting objects (in this research, they refer to the pornographic web pages), namely the proportion of the pornographic web pages that are accurately classified as pornographic. The higher this index, the greater the filtering capabilities of the method and the lower the rate that pornographic web pages are judged as non-pornographic. This index is calculated through the following formula:

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

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:

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

For example, when the value of precision is too high while the value of recall is too low, it means that although the pornographic websites could be filtered, the chances of non-pornographic ones being misjudged are also very high. 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 [33]. 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 4. Note that the training and unknown web pages should be selected randomly from the three categories.

Table 4. The numbers of web pages used in each phase

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

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 5, 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 6, the proportion of pornographic web pages that were accurately filtered was (TP) 97.33%, while the proportion of non-pornographic web pages that were accurately filtered was (TN) 87.52%.

Table 5. 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 |

Table 6. The efficiency of classification without using the Medical Keyword Table

| Indexes | Measurement | Indexes | Measurement |
|---|---|---|---|
| TP | 97.33% | Accuracy | 90.79% |
| TN | 87.52% | Precision | 97.04% |
| FP | 2.67% | Recall | 87.52% |
| FN | 12.48% | F-measure | 83.01% |

Then, we perform the Classification Phase by applying the Medical Keyword Table. As shown in Table 7, 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 8 recorded the classification efficiency of this experiment. By comparing Table 8 with Table 6, we observed that all the efficacy assessment indexes of Accuracy, Precision, Recall, and F-measure were improved noticeably. Moreover, FN 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.

Table 7. 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 8. The classification efficiency of using the Medical Keyword Table

| Indexes | Measurement | Indexes | Measurement |
|---|---|---|---|
| TP | 97.33% | Accuracy | 96.67% |
| TN | 96.33% | Precision | 97.31% |
| FP | 2.67% | Recall | 96.34% |
| FN | 3.67% | F-measure | 95.86% |

**(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.

Table 9. 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 10. The effectiveness of the Relearning Phase

| Indexes | Measurement | | Indexes | Measurement | |
|---|---|---|---|---|---|
| | Case (I) | Case (II) | | Case (I) | Case (II) |
| TP | 97.95% | 97.99% | Accuracy | 96.21% | 98.26% |
| TN | 95.32% | 98.36% | Precision | 97.89% | 98.00% |
| FP | 2.05% | 2.01% | Recall | 95.32% | 98.36% |
| FN | 4.68% | 1.64% | F-measure | 94.06% | 98.54% |

The experimental results were shown in Table 9 and 10. 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 9, 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 10. By using the Relearning Phase, the evaluation indicator FN (the rate of non-pornographic web pages being misjudged as pornographic web pages) decreased from 4.68% to 1.64%. Moreover, the Accuracy increased from 96.21% to 98.26% while the Precision increased from 97.89% to 98.00%. This means that both the Accuracy and Precision were improved after the Relearning Phase was turned on; after the re-learning, TP (the rate of pornographic web pages being accurately judged as pornographic web pages) increased from 97.95% to 97.99% while FP (the rate of pornographic web pages being judged as non-pornographic web pages) decreased from 2.05% to 2.01%, showing a slight improvement in terms of the filtering performance concerning pornographic web pages. TN (the rate of non-pornographic web pages classified as non-pornographic web pages) increased from 95.32% to 98.36%, a significant increase in terms of the classification of normal web pages. FN 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 the stability of the filtering mechanism adopted in the current research. Table 11 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 web pages, 200 pornographic web pages, and 300 medical web pages.

Table 11. The experimental results of six data groups under various combination ratios

| Group No. | Total number of web pages | Ratio | Accuracy (%) | Precision (%) | FP (%) | FN (%) |
|---|---|---|---|---|---|---|
| **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 |

Obviously, some changes occurred over the four measuring indicators of Accuracy, Precision, FP and FN, though not very substantial changes; when medical web pages accounted for a higher proportion, the FN (the proportion of non-pornographic pages being misjudged as pornographic web pages) 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 FN were fair proof that the method in this research was satisfactory.

# 6. 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 page files and

establishing decision trees according to the category of web pages. The association rules in each decision tree are thus generated and scored. The rule scores are used ultimately 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 FN, 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. The third experiment was to test the stability of the proposed method. The experimental results indicated that the filtering results of our method would not be greatly influenced by the changes in the data. 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.

## Reference


[1] A.Ahmadi, F.Fotouhi, M.Khaleghi. Intelligent classification of web pages using contextual and visual features. Applied Soft Computing,2010.

[2] Bernadette H. Schell and Clemens Martin, Cybercrime A Reference Handbook, Oct. 2004.

[3] B. Stayrynkevitch, et al., Poesia software architecture definition document, in: Technical Report, Poesia Consortium, Dec., 2002.

[4] CAPHIS/MLA., The Librarian’s Role in the Provision of Consumer Health Information and Patient Education. Bulletin of the Medical Library Association. vol. 84, pp. 238, 1996.

[5] C. N.Wathen, R. M.Harris, An examination of the health information seekingexperiences of women in rural Ontario,Canada. Information Research, vol. 11, no. 4, pp. 267, 2006.

[6] Chen-Huei Chou, Atish P. Sinha, Huimin Zhao, “Commercial Internet filters: Perils and opportunities ,”Decision Support Systems, vol. 48, no 4, Mar. 2010, pp. 521–530.

[7] Edward A. Cavazos, “Cyberspaceand the Law : Your Rights and Duties in the On-line World,”Cambridge, Mass. ; London, England : MIT Press, 1994, pp. 90-93.

[8] F.A.Sonnenberg, Health Information on the Internet: Opportunities and Pitfalls.Archives of Internet Medicine, vol. 157, no. 2, pp. 151-152, 1997.

[9] J. A.Sangl, L. F.Wolf, Role of consumer information in today's health care system. Health Care Financing Review, vol.18, no. 1, pp. 1-8, 1996.

[10] J. AHartigan, Clustering algorithms. Wiley, New York,1975.

[11] J.Liu, X. Li, and W. Zhong, “Ambiguous decision trees for mining concept-drifting data streams,” Pattern Recognition Letters, vol.33, pp. 1347-1355, 2009.

[12] J.R. Quinlan, “Induction of decision trees,” Machine Learning, vol. 1, pp. 81-106, 1986.

[13] J.R.Quinlan, “C4.5:Programs for Machine Learning,” San Mateo: Morgan Kaufmann, 1993.

[14] K.Mitchell, D.Finkelhor, J.Wolak,“The exposure of youth to unwantedsexual material on the Internet: A national survey of risk, impact, andprevention,” YOUTH & SOCIETY, vol. 34 no. 3, Mar. 2003, pp. 330–358.

[15] L. Breiman, H. J. Friedman, R. A. Olshen, and C. J. stone, “Classification and regression trees,” Belmont, Wadsworth International Group, 1984.

[16] L. H. Lee, and C. J. Luh, “Generation of pornographic blacklist and its incremental update using an inverse chi-square based method,” Information Processing and Management, vol. 44, pp.1698-1976, 2008.

[17] L. Y. Michele, F. David, J. M. Kimberly, and W. Janis, “Associations between blocking, monitoring, and filtering software on the home computer and youth-reported unwanted exposure to sexual material online,” Child Abuse & Neglect, vol. 33, pp. 857-869, 2009.

[18] M. Hammami, Y. Chahir, and L. Chen, “ WebGuard: A Web Filtering Engine Combining Textual, Structural, and Visual Content-Based Analysis,” IEEE Transactions on Data Engineering, vol. 18, pp. 272-284, 2006.

[19] M.M.Fleck, D.A.Forsyth and C.Bregler. Finding naked people, Proc.European Conf. on Computer Vision,1996.

[20] P. Y. Lee, S. C. Hui, and A. C. M. Fong, “An Intelligent Categorization Engine for Bilingual Web Content Filtering,” IEEE, vol. 7, pp. 1183-1190, 2005.

[21] W. Ho, P. Watters, Statistical and structural approaches to filtering internetpornography, in: Proceedings of the IEEE International Conference on Systems,Man and Cybernetics, pp. pp. 4792–4798, 2004.

[22] Z. Zhang, and O. Nasraoui, “Mining search engine query logs for social filtering-based query recommendation,”Applied Soft Computing, vol. 8, no 4, Sep. 2008, pp. 1326–1334.

[23] CyberSitter, http://www.cybersitter.com/

[24] ICRA, Internet Content Rating Association, http://www.fosi.org/icra/

[25] Internetworldstats, http://www.internetworldstats.com/stats.htm

[26] MedlinePlus, http://www.nlm.nih.gov/medlineplus/

[27] Online MBA, http://www.onlinemba.com/online-mba/

[28] Pew Internet, http://pewinternet.org/

[29] PICS, Platform for Internet Content Selection, http://www.w3.org/PICS

[30] Planned Parenthood, http://www.plannedparenthood.org/

[31] SafeSquid® HTTP FIREWALL, http://www.safesquid.com/

[32] TopTenREVIEWS Expert Product Reviews, http://www.toptenreviews.com/

[33] URL blacklist service, http://urlblacklist.com/

[34] W3schools, http://www.w3schools.com/

[35] W3C, World Wide Web Consortium, http://www.w3.org

# 科技部補助計畫衍生研發成果推廣資料表

日期:2015/10/30

| | |
|---|---|
| 科技部補助計畫 | 計畫名稱: 一個能兼具相似度與差異度計算以及再學習機制的有效率電子文件辨識方法：以色情及醫學網頁辨識為例 |
| | 計畫主持人: 許志堅 |
| | 計畫編號: 103-2410-H-004-112- 學門領域: 資訊管理 |

無研發成果推廣資料

# 103年度專題研究計畫研究成果彙整表

| | |
|---|---|
| **計畫主持人：**許志堅 | **計畫編號：**103-2410-H-004-112- |
| **計畫名稱：**一個能兼具相似度與差異度計算以及再學習機制的有效率電子文件辨識方法：以色情及醫學網頁辨識為例 | |

| 成果項目 | | | 量化 | | | 單位 | 備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等） |
|---|---|---|---|---|---|---|---|
| | | | 實際已達成數（被接受或已發表） | 預期總達成數（含實際已達成數） | 本計畫實際貢獻百分比 | | |
| 國內 | 論文著作 | 期刊論文 | 0 | 0 | 100% | 篇 | |
| | | 研究報告/技術報告 | 0 | 0 | 100% | | |
| | | 研討會論文 | 0 | 0 | 100% | | |
| | | 專書 | 0 | 0 | 100% | 章/本 | |
| | 專利 | 申請中件數 | 0 | 0 | 100% | 件 | |
| | | 已獲得件數 | 0 | 0 | 100% | | |
| | 技術移轉 | 件數 | 0 | 0 | 100% | 件 | |
| | | 權利金 | 0 | 0 | 100% | 千元 | |
| | 參與計畫人力（本國籍） | 碩士生 | 0 | 2 | 100% | 人次 | |
| | | 博士生 | 0 | 0 | 100% | | |
| | | 博士後研究員 | 0 | 0 | 100% | | |
| | | 專任助理 | 0 | 0 | 100% | | |
| 國外 | 論文著作 | 期刊論文 | 0 | 3 | 100% | 篇 | |
| | | 研究報告/技術報告 | 0 | 0 | 100% | | |
| | | 研討會論文 | 0 | 0 | 100% | | |
| | | 專書 | 0 | 0 | 100% | 章/本 | |
| | 專利 | 申請中件數 | 0 | 0 | 100% | 件 | |
| | | 已獲得件數 | 0 | 0 | 100% | | |
| | 技術移轉 | 件數 | 0 | 0 | 100% | 件 | |
| | | 權利金 | 0 | 0 | 100% | 千元 | |
| | 參與計畫人力（外國籍） | 碩士生 | 0 | 0 | 100% | 人次 | |
| | | 博士生 | 0 | 0 | 100% | | |
| | | 博士後研究員 | 0 | 0 | 100% | | |
| | | 專任助理 | 0 | 0 | 100% | | |

| | |
|---|---|
| 其他成果（無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文 | 本計畫主要成果在於撰寫為國際SCI期刊論文：<br>目前已經投稿了兩篇SCI期刊、另有兩篇SCI期刊稿件正在撰寫之中(上表中無法紀錄此狀況)。 |

| 字敘述填列。） | |
|---|---|

| | 成果項目 | 量化 | 名稱或內容性質簡述 |
|---|---|---|---|
| 科教處計畫加填項目 | 測驗工具(含質性與量性) | 0 | |
| | 課程/模組 | 0 | |
| | 電腦及網路系統或工具 | 0 | |
| | 教材 | 0 | |
| | 舉辦之活動/競賽 | 0 | |
| | 研討會/工作坊 | 0 | |
| | 電子報、網站 | 0 | |
| | 計畫成果推廣之參與（閱聽）人數 | 0 | |

# 科技部補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估
   ■達成目標
   □未達成目標（請說明，以100字為限）
   　□實驗失敗
   　□因故實驗中斷
   　□其他原因
   說明：

2. 研究成果在學術期刊發表或申請專利等情形：
   論文：□已發表 □未發表之文稿 ■撰寫中 □無
   專利：□已獲得 □申請中 ■無
   技轉：□已技轉 □洽談中 ■無
   其他：（以100字為限）

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）（以500字為限）

   ・在學術研究方面：
   本計劃以知識學習的觀念，運用資料探勘技術中的決策樹演算法來研究樣本網頁，找尋網頁文件欄位之間特有的關連性規則；並且依據色情網頁與醫學網頁之間相似與相異的特徵進行區分，在提升對於色情網頁之過濾能力的同時，也能避免造成醫學網頁被誤判。因此本計劃的研究成果可望更深入地學習網頁辨識相關的知識，並且發現決策樹資料探勘方法在處理網頁過濾問題時的執行條件的設定以及運用上的相關現象，可作為引進其他資料探勘方法的參考，從理論面尋找可能的更好作法。

   ・在產業與社會責任方面：
   色情網站的問題是目前網路安全、網路管理上的一個重大課題，它嚴重地影響青少年、兒童的身心發展，造成極大的社會問題，近年來已經嚴重地戕害兒童、青少年身心並且引發一連串驚世駭俗的治安事件，如果能提供一個有效的過濾色情網頁機制演算法，亦將是學術研究從業者對於社會的一種回饋。此外，其他領域的電子文件分類也是一個熱門的系統發展方向，藉由本計畫之研究，將可提供業者作為未來改良系統效率與準確性之參考。

   ・在研究論文產出方面：
   計畫主持人近年已在資料探勘相關題目進行相關研究並獲得發表成果，本計畫

之研究成果也已經進行彙整與擴充，並撰寫成為英文論文，預計將撰寫二至三篇相關論文，投稿至國際SCI期刊。