運用以卡方為基礎的統計方法於色情網頁分類之研究

李龍豪
元智大學資訊管理學系

陸承志
元智大學資訊管理學系

摘要

由於網際網路的普及，資訊的散佈非常迅速，網路上充斥着各種良莠不齊的資訊，越來越多的不當資訊，例如色情小說、圖片與粗暴文字等，在缺乏完善的網路內容管理機制之下，使用者只要透過搜尋引擎輸入相關的關鍵字，就可以從搜尋結果藉由超連結輕易存取網站內容，因此網路內容管理已成為刻不容緩的議題。本研究針對不當資訊中的色情範疇，提出一個以色情網頁分類，來蒐集黑名單的方式，對色情網站內容中文字的部份，求出個別字詞(Word)的色情傾向(Porn Tendency)，透過卡方分配計算出色情指標值(Indicator Value)，將網頁分成色情(Porn)、未確定(Unsure)與非色情(Non-Porn)三類。色情類網頁的網址即為所謂的黑名單，可做為網路色情過濾的依據。本研究針對中文與英文語系網頁實作一個系統，實驗結果顯示，本提議方法具有高度的精確率與相當低的正誤判率。

關鍵詞：網路內容分類、色情黑名單、不當資訊過濾、卡方分配。

Classifying Pornographic Web Pages Using a Chi-Square Based Statistics Method

Lung-Hao Lee
Department of Information Management, Yuan Ze University

Cheng-Jye Luh
Department of Information Management, Yuan Ze University

Abstract

With the rapid growing of Internet usage, inappropriate materials (e.g. porn, drug, violence et al.) had been flooded on the Web. The open characteristic of the Web allows users to access almost any type of such inappropriate materials, consequently having various negative effects on the users, particularly on the children. Thus, web content rating and filtering mechanism is a worthy and pressing issue. This study proposes a chi-square based statistics method for classifying pornographic materials. Given a web page, its textual content is first split into a list of tokens, along a porn tendency weight for each token. The proposed method then calculates an indicator value (I-value) for the web page by combining the tokens' porn tendency weights through properties of chi-square distribution. The resulting I-value is used to classify the given web page into one of three categories, Porn, Unsure and Non-Porn. The web pages in the Porn Category are finally collected into a black list. Currently, the proposed method can classify English and Chinese Web pages. Experimental results indicate that the proposed method can detect pornographic web content at a superior precision rate along with a very low false positive rate.

Key words: Web Content Rating, Pornographic Black List, Inappropriate Web Content Filtering, Chi-Square Distribution

壹、緒論

網際網路的蓬勃發展，使得資訊的流通非常迅速與廣泛。面對不斷產生的多樣化資訊，如何針對網路內容做適當的分級與管理 (Balkin et. al. 1999; Goodwin & Vidgen 2002) 一直是一個被廣泛討論的議題，尤其是不當資訊的偵測與防治更受到特別關注。不當的網路資訊包含色情、暴力、吸毒、賭博等領域，青少年及兒童在養成教育時接受到不當資訊的影響，容易戕害身心的健全發展，導致人格的偏差，造成眾多社會與犯罪問題。

在網路分級方面，W3C 在西元 1997 年提出 Platform for Internet Content Selection (PICS 1997-2003)，做為網頁內容分級的標準規格，內容提供者可以依此規格制定的標籤(labels)對內容自我分級，或是由協力廠商對網路上散佈的內容做分級。由於業者的商業利益考量，由網路內容提供者自律的做法成效不彰，根據新加坡學者 Lee et al. (2002&2003) 的調查，網路內容採用 PICS 者大約僅佔 11.0%。所以由網路頻寬提供者，例如網路頻寬業者、政府、機關團體等，來做偵測與防範似乎是目前較為可行的方式。

我國自 2005 年 10 月起開始施行網站內容分級管理辦法，該辦法要求網路內容提供者將網站內容對照國際 The Internet Content Rating Association (ICRA 1999-2006；林宜隆等人民 92) 所制定詞彙標準，將網路內容依屬性區分為語言、性與裸露、暴力及其他等分類，再依分類內容區分為限制級、輔導級、保護級、普通級四種。例如在語言上出現明顯與性相關字眼，即分類為限制級。

我們認為實施網站內容分級之前，最重要的工作是要能將限制級的網站內容（例如色情、暴力、毒品等）與其他非限制級的內容做適當的區分，然後再進行其他級別的篩選。同時，我們認為限制級的網站內容可以透過資訊技術來篩選，其他級別（尤其是輔導級、保護級）多少具有自由心證，短期內難以透過資訊技術來區分。

本研究針對不當網路內容中最為氾濫的網路色情，提出一個以卡方為基礎的色情網頁偵測的方法。此一方法對每一個網頁中的文字部分，運用統計推論中的卡方分配特性，計算出一個介於 0 到 1 之間的色情指標值(Indicator Value, I-value)，再依臨界值(threshold)將 I 值對應到三個互斥區間的其中一個區間，以判定該網頁是屬於色情(Porn)、未確定(Unsure) 或非色情(Non-Porn)。實驗結果顯示，本研究所提出的色情網頁偵測方法具有高度的精確率與低度的正誤判率。

本文其他章節安排如下: 第二節進行本研究的相關文獻探討，包括網路探勘、色情分類的相關研究以及卡方文字分析方法；第三節提出一個卡方為基礎的色情網頁分類方法；第四節說明本研究的系統實作與實驗評估；最後，第五節總結本研究的成果與貢獻，並闡述未來的研究方向。

貳、文獻探討

一、網路探勘 (Web Mining)

網路探勘是一種應用資料探勘技術從網頁資料中萃取知識的研究。Etzioni 在 1996 年率先提出這個名詞，並且探討網頁是否足夠結構化，以便有效率的進行資料探勘，開啟了網路探勘這個研究領域的濫觴。許多學者陸續針對這個研究領域提出相關的分類與探討 (Jicheng et al. 1999 ; Kosala & Blockeel 2000 ; Liu et al. 2002 ; Srivastava et al. 2002 ; Kolariand & Joshi 2004)。這個研究領域 (如圖 1 所示) 大致上可以分成網路內容探勘 (Web Content Mining, WCM)，網路結構探勘 (Web Structure Mining, WSM) 與網路瀏覽探勘 (Web Usage Mining, WUM) 三個子領域。其中，「網路內容探勘」針對個別網頁的內容，包含文字、圖片、聲音與影像等等進行資料探勘；「網路結構探勘」探討個別網頁的文件結構，或者是網頁之間的超連結結構的組成關係，依照鏈結的不同又可分為網站內連結與網站間連結；「網路瀏覽探勘」利用資料探勘的技術從記錄檔中挖掘使用者行為關係。

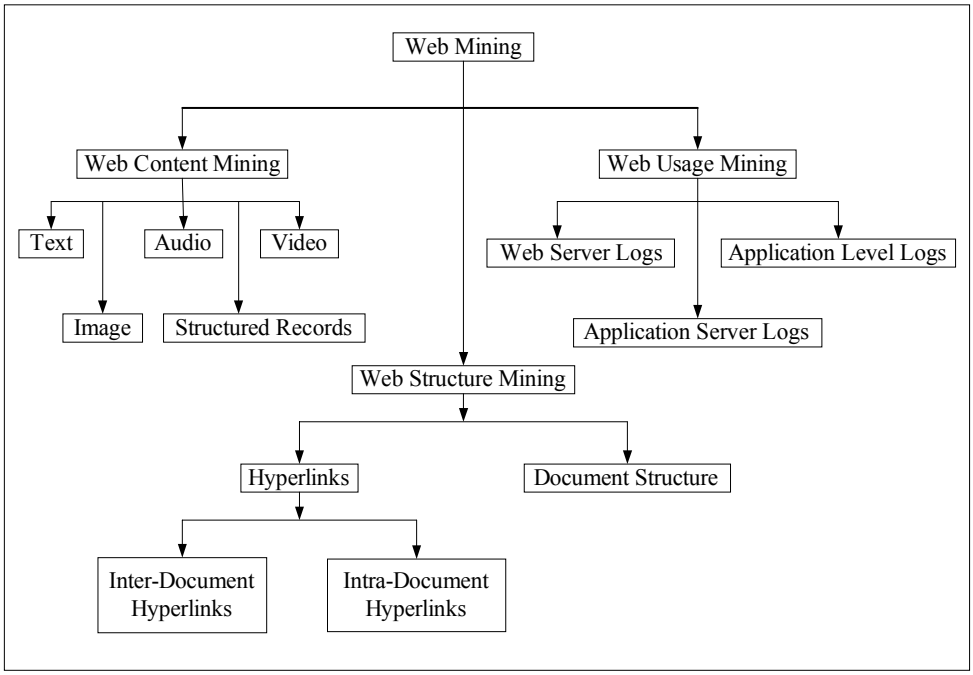

圖 1 網路探勘研究領域的分類 (修改自 Srivastava et al. (2002))

本研究所提出的方法，針對網頁中的文字部分做統計運算，求出每一個網頁的色情指標值，用來鑑別網頁內容中的色情成分，屬於網路內容探勘的研究。其次，我們會利用網頁對內及對外的連結來蒐集相關聯的網頁，這部分屬於網路結構探勘。相對

的，許多針對網站記錄檔做資料探勘以辨別色情網路內容的研究 (楊良吉 民 90；邱志傑等人 民 92；王鐵雄等人 民 93)，就屬於網路流覽探勘。以下我們探討不同的色情網頁分類方法，並將相關方法的優缺點整理如表 1。

二、色情網頁分類

（一）記錄檔分析方法(Log-based Approach)

這個方法不管網站內容與表現形式，主要針對 proxy server 或 web server 的記錄檔做資料分群，以取得色情黑名單 (楊良吉 民 90)。邱志傑等人 (民 92)分析來自包含台大、交大、中興、中山與成大等五個區域中心之 14 台 Proxy 的瀏覽記錄，並在國立成功大學計算機中心建立不當資訊防治系統(http://tanet110.ncku.edu.tw/)，用來過濾不當色情網頁。王鐵雄等人(民 93)則從社會心理學的從眾行為觀點，從網站記錄檔中分析出使用者最常瀏覽的前 300 筆不當資訊色情網站，試圖以最少的黑名單達到最大的過濾效果，並以台灣學術網路現行階層式管理架構為基礎，提出一套區域聯防的機制。本方法的優點在於若過濾主題的使用者行為較為明顯且具有一定的特性 (例如多人重覆瀏覽同一網頁)時，可以獲得較高的精確率，然而當使用者行為不明顯或是改變迅速時，就難以從記錄檔中透過資料探勘的技術得到預期的效果。

（二）內容分析方法(Content-based Approach)

根據網路色情資訊種類不同，內容分析方法又可以區分為三種方式:

1. 純文字分析

目前過濾相關議題最常採用純文字分析方式，因為文字資訊量小且為網站內容提供者最容易使用的表現形式，而且對文字做分析的複雜度低，同時可以獲得較好的效能。Lee et al. (2002&2003) 從訓練網頁的文字內容中找出 55 個常用的色情關鍵字來建立網頁的特徵向量 (term frequency vector)，再透過類神經網路計算關鍵字的權重(weight)，然後對測試網頁做分群，以判別網頁是否為色情類別。該研究宣稱其精確度可達 95.0%。邱忠俊(民 88)則是從犯罪語言學的角度，搜集與分析網路情色文學內容後，建立一個情色語料庫。該研究在透過設計的資訊檢索模型下，測試 1164 筆網頁的平均精確率為 85.67%。

2. 圖片分析

本方法對色情圖片做影像處理，找出圖片中可能與人體皮膚區塊有關的部份，描繪出皮膚的統計特徵，再經由大批圖片的訓練找到色情圖片的型樣。在判斷某個網頁是否為色情時，本方法先將單一圖片的像素特徵與色情型樣做型樣識別來判斷該圖片是否為色情圖片，然後根據網頁中色情圖片的數量比率來判別是否為色情類別 (Chan et al. 1999；Smith et al. 1999；Jiao et al. 2001；Bosson et al. 2002；郭永明 民 90) 。在相關實作部分，Duan et al. (2002) 使用皮膚顏色模型搭配 Support Vector Machine (SVM) 偵測成人影像的精確率為 87.3%。Schettini et al. (2003) 結合 Classification and

Regression Trees (CART) 以及 SVM 建構出一個色情決策樹，在色情與非色情圖片總數共 1500 張的實驗下，作者宣稱其平均精確率為 89.4%。 Arentz & Olstad (2004) 則不同於一般分析色情圖片建構特徵向量的方式，對於任一張影像給定某個機率值，用來表示該張圖片包含情色內容的可能性，再透過基因演算法來演化計算，藉此建構色情特徵向量，實驗結果顯示，該方法的精確率為 89%。

圖片分析常見的問題是對大頭照的誤判，因為大頭照的膚色佔整張圖片比率過高，常被誤判為色情圖片。至於網路色情的動態影像部分，數量上比文字與圖片相對的少很多，且通常出現在收費網站或是私人 FTP 站中，加上通常使用視訊串流的技術，無法藉由搜尋引擎搜尋輕易取得，可能的作法是將畫面自影像中擷取出來，然後再做色情圖片分析 (Torres & Vila 2002)。另外，由於影像處理比文字分析要花費更多的系統資源與運算時間，較難滿足線上即時過濾的效能需求。

3. 圖文綜合分析

Hammami et al. (2003) 提出 Web Guard 系統架構，結合文字內容、圖片內容與 URL 建構出特徵向量，再利用資料探勘技術將網頁區分為正常網頁 (Normal URLs) 與嫌疑網頁 (Suspect URLs) ，該方法實驗的精確度為 95.0%。邱志傑等人(民 93)利用人工過濾的方式挑出 169 個中文不當關鍵字詞，並定義色情影像為圖片中裸露女性胸前特徵，提出一個胸前特徵偵測演算法，試圖改善文字分析的不足。該研究混合影像測試結果顯示，判別精確度為 84.51%。邱建明(民 93)提出一個結合文字與影像的色情偵測方法，將每一筆分析的網頁用 5 個特徵去表示，利用這些特徵向量與測量相似性的方式建出相似矩陣。接著，依據網頁之間的關係建構 Web Graph，其中邊上的權重則是兩兩文件的相似度，再利用圖形切割 (Graph Partitioning)的方法找出色情網頁在圖形中的子圖。該方法實驗結果精確度為 88.0%。

表 1 色情網頁分類方法比較

作者	分析方法	優點	缺點
楊良吉 (民 90)	針對 proxy or web server 紀錄檔做資料分群。	(1) 使用者行為較為明顯且具有一定的特性時，可以獲得較高的精確率。 (2) 計算時間較少。	當使用者行為不明顯或是改變迅速時，無法有效判斷。
邱志傑等人 (民 92)	針對 5 個區網中心的 proxy 中的瀏覽資料做分析。		
王鐵雄等人 (民 93)	從社會心理學的從眾行為觀點分析記錄檔。		
Lee et al. (2002&2003)	(1) 純文字分析方法。 (2) 類神經網路分析。	(1) 文字資訊量小。 (2) 純文字分析的複雜度低。	須具備背景知識。
邱忠俊 (民 88)	(1) 純文字分析方法。 (2) 建立一個情色語料庫。		

Duan et al. (2002)	(1) 圖片分析方法。 (2) 使用皮膚顏色模型搭配 Support Vector Machine (SVM) 偵測成人影像。	(1) 圖片分析方式沒有語言識別的問題，一律都是皮膚區塊的型樣識別。	(1) 影像處理比文字分析要花費更多的系統資源與運算時間。 (2) 大頭照等皮膚比率較高的圖片，容易造成誤判。
Schettini et al. (2003)	(1) 圖片分析方法。 (2) 結合 Classification and Regression Trees (CART) 以及 SVM 建構出色情決策樹。		
Arentz & Olstad (2004)	(1) 圖片分析方法。 (2) 利用基因演算法演化與調整色情特徵向量。		
Hammami et al. (2003)	(1) 圖文綜合分析方法。 (2) 提出 Web Guard 系統架構，結合文字內容、圖片內容與 URL 建構出特徵向量。	針對網頁的內容(文字與圖片)全面分析。	處理時間較長，耗費相對較多的系統資源。
邱志傑等人 (民 93)	(1) 圖文綜合分析方法。 (2) 人工挑選 169 個中文不當關鍵字詞。 (3) 提出一個胸前特徵偵測演算法。		
邱建明 (民 93)	(1) 圖文綜合分析方法。 (2) 以網頁的特徵向量與測量相似性的方式建出相似矩陣。 (3) 建構網頁的 Web Graph。 (4) 利用 Graph Partitioning 找出色情網頁在圖形中的子圖。		

三、卡方文字分析

本研究採用的卡方分析(Chi-Square Statistics)方法參考自垃圾郵件防堵(anti-spamming) 相關研究，底下針對這些研究進行探討。

首先，Graham (2002) 提出一個貝氏垃圾郵件過濾器，這個過濾器在訓練階段時，針對訓練用的正常郵件 (ham) 與垃圾郵件 (spam)，計算出每一封郵件中每個字詞的出現頻率(word frequency)，然後在測試階段，就根據測試郵件中出現的所有特定字詞，以貝氏統計演算法算出這些個別字詞結合出現的機率(combining probability)，求出該郵件為垃圾郵件的可能性。該方法的實驗結果顯示: 正常郵件與垃圾郵件兩條分佈曲線的走向極類似，接近於重疊，沒有顯著的差異可以達到偵測垃圾郵件的效果。

接著，Robinson 試圖改善 Graham 方法的缺失，初步嘗試結合個別字詞，出現的機率求出每一封郵件對應的分數，雖然這個方法可以大致將成正常郵件與垃圾郵件區分成兩大塊，但中間部分卻有一個明顯重疊的區段，造成選擇割裂點 (cutoff) 的困

難。接著，Robinson 再次嘗試利用中央極限定理(Central Limit Theorem)(Ross 2004) 計算個別字詞的結合機率，雖然可以有效區分正常郵件與垃圾郵件，但會產生一塊令人混淆的不確定(Unsure)區域 (Anthony, 2003)。之後，Robinson (2003) 提出一個 Chi-Square based 字詞結合方法，實作測試的結果明顯地可以看到正常郵件與垃圾郵件會往兩邊集中，中間只剩極小的不確定部分，這部分是因為該郵件沒有顯著的證據顯示為正常郵件或是垃圾郵件。

不過，在許多引用 Robinson (2003) Chi-Square based 字詞結合方法在垃圾郵件偵測的研究發現，用來判斷的垃圾郵件傾向值會受到結合的字詞數目影響，導致該方法無法達成有效區別垃圾郵件的效果。於是，有學者提出 Effective Size Factor (ESF) 的概念來調整結合的字詞數目(Robinson, 2004; Meyer & Whateley, 2004)，但目前尚處於測試驗證階段。目前採用 Robinson (2003) 方法的垃圾郵件過濾計劃有 SpamBayes (http://spambayes.sourceforge.net/) 與 Bogofilter (http://bogofilter.sourceforge.net/)。

參、卡方色情網頁分類

一、系統流程

整個蒐集色情黑名單的流程如圖 2 所示。首先，我們使用 Web Crawler 從網際網路上抓取跟色情關鍵字相關的網頁，再從其中選取部份網頁做為訓練資料，經由訓練階段產生 Token Database，其餘網頁則做為測試資料。本研究利用一個統計方法的卡方分配特性，對每個測試網頁求出一個色情指標值(Indicator Value, I-Value)，以根據此一色情指標值將網頁區分為色情(Porn)、未確定(Unsure)以及非色情(Non-Porn)三個種類，再將其中的色情網頁加入色情黑名單(URL Black List)中。底下我們詳細說明每個步驟。

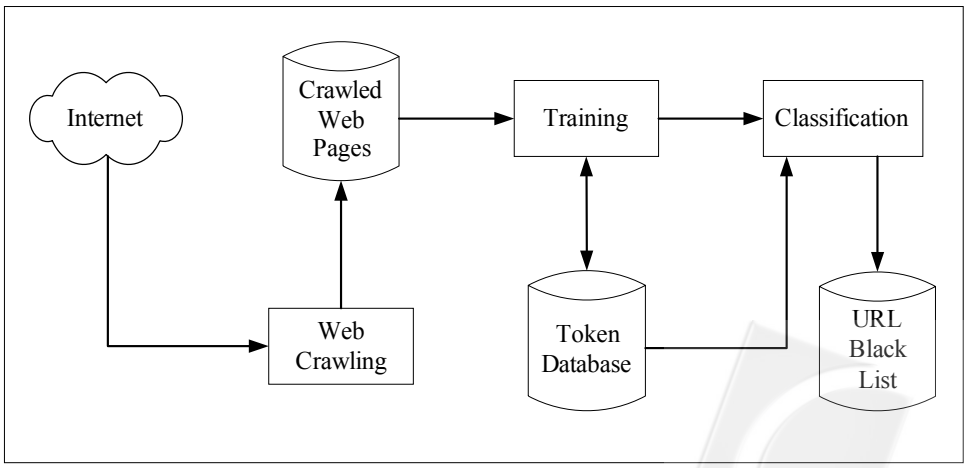

圖 2 蒐集網路色情黑名單流程圖

二、網頁擷取 (Web Crawling)

Web Crawling 步驟如圖 3 所示。首先，我們以程式自動對搜尋引擎輸入預先蒐集的色情關鍵字，例如：色情小說，情色貼圖等，然後將搜尋引擎回傳的搜尋結果中的網址萃取出來當作種子網址(Seed URLs)，以做為產生網頁資料集的起始點。然後，我們將每個種子網址的網頁內容抓取回來，再將這些抓取回來網頁中的內部連結(Embedded URLs)萃取出來，執行下一輪的擷取動作。由於色情網站有彼此透過外部超連結緊密互連的特性，透過重複以上的萃取網址與抓取網頁的動作，可以蒐集較多潛在的色情網頁。

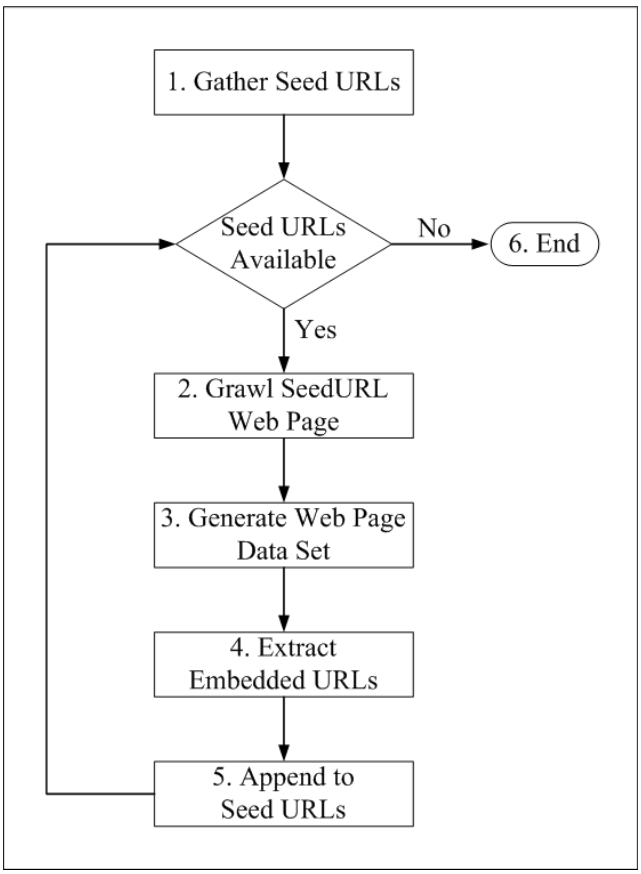

圖 3　網頁擷取步驟

三、訓練(Training)

我們從蒐集回來的網頁資料中，以人工方式選取部份的網頁做為訓練資料。訓練資料的界定由我們實驗室內三名研究生以人工方式篩選，只要至少有兩人認為是色情的即是色情網頁；另一方面，只要至少有兩人認為是非色情的即是非色情網頁；其他

的則是“未確定”。由於研究生的年紀、背景相仿，所以在色情與非色情的判定上相當一致。“未確定”的網頁則通常含有具有爭議的網頁內容，例如性別教育議題，或是討論區等形式的網頁，這些內容在抽樣調查下,大約有相當的人數判定為色情或非色情,或者有大多數人認為無法判定。其中色情部分的訓練資料，我們盡可能選擇不同形式的網頁內容 (Lee et al. 2002 & 2003；邱忠俊 民 88)，例如以文字呈現為主的色情小說，或者是以色情寫真圖片為主的色情網站，以反應網際網路上色情網站實際情形。同樣的在非色情這個部份，除了多樣式的網頁內容之外，也應該包含健康、醫療、反色情、性知識等等議題的內容，因為這部份的網頁內容常會出現色情關鍵字，例如：人體器官名稱、性愛動作等，如果非色情訓練部分沒有包含這些議題的資料，訓練過後產生的 token 的色情傾向值(porn tendency)會有誤差，導致這些相關議題的網頁會在測試階段被誤判為色情網頁。在訓練時，我們只採用色情與非色情兩類網頁，以有效訓練系統對色情的鑑別度。在實測時，系統所建議的未確定網頁表示系統無法鑑別該網頁是色情或非色情。訓練過程的詳細步驟如圖 4 所示。

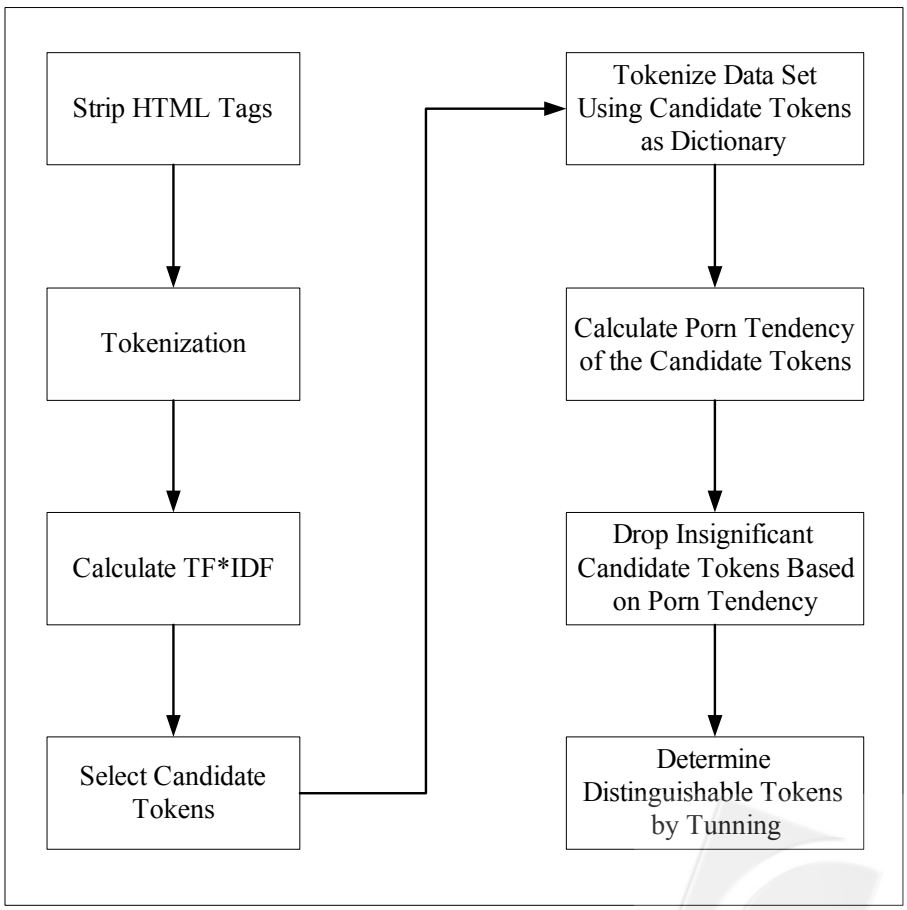

圖 4 訓練流程

（一）Strip HTML Tags

首先，我們移除訓練資料中的網頁標籤，只留下在<title>、<meta>、<body>內的網頁文字內容。在去除標籤過後，本研究選擇資訊量大小(以位元組計)相當的色情與非色情文字內容進行底下的訓練步驟。

（二）Tokenization

此步驟將上述兩部份的文字內容做斷詞處理，即將原先可能是整篇的文章、段落或句子，分成一個個具有語意的字詞。

（三）Calculate TF*IDF

我們採用資訊檢索(Information Retrieval)領域中常用的特徵表示方式，對每個字詞計算 TF*IDF (Baeza-Yates & Ribeiro-Neto 1999)。TF*IDF 值越大表示這個字詞出現的廣度與深度達到較佳的狀態，也就是該字詞出現在較多的網頁中，且在該網頁中出現的頻率較高。

（四）Select Candidate Tokens

我們將字詞的 TF*IDF 值由大到小做遞減排序，選擇前面一定數量的字詞，例如：TF*IDF 較大的 1000 個字詞，做為 Candidate Tokens。

（五）Tokenize Data Set Using Candidate Tokens as Dictionary

此步驟將色情與非色情兩部分的 Candidate Tokens 混合成一個字典，再對原先的訓練資料做一次斷詞處理，找出這些 Candidate Tokens 在色情與非色情部分的訓練資料中出現的情形。

（六）Calculate Porn Tendency of the Candidate Tokens

然後，我們算出色情部分 Candidate Tokens 的色情傾向值(Porn Tendency)，計算方式為相對比率的概念，即將色情部份個別 Candidate Token 的 TF*IDF，除以色情部份個別 Candidate Token 的 TF*IDF 加上非色情部份個別 Candidate Token 的 TF*IDF 的總和。舉例來說，假設色情部分有一個 Candidate Token “情色”，計算過後得到的 TF*IDF 值為 700，而“情色”這個字詞在非色情部份的 TF*IDF 是 300，則“情色”這個字的色情傾向值為 700 除以 1000 等於 0.7。

（七）Drop Insignificant Candidate Tokens Based on Porn Tendency

我們將色情傾向值由大到小排列，留下最前面百分之二十五與最後面百分之二十五數量的字詞，移除中間百分之五十的字詞。色情傾向值越高表示這個字詞出現在色情網頁的可能性越大；相對地，色情傾向值越低的字詞出現在色情網頁的可能性較低；而色情傾向值位於中間的字詞則不具代表性，例如：假設“色情”這個字詞的色情傾向值為 0.6，排序之後位於整體的中間的部分，表示“色情”這個字詞可能出現在色情網站的可能性比非色情來得高，但並沒有明顯的差異，例如在某些反色情的網站，或是色

情相關問題的新聞報導中也會出現"色情"這個字詞，所以這個字詞不能拿來有效地區分色情與非色情，所以我們應該移除沒有顯著色情傾向的字詞。

（八）Determine Distinctive Tokens by Tuning

最後，這個步驟是用來調校精確率。理論上，經由上述 7 個步驟挑選出來的 Candidate Tokens 具有識別色情與非色情的鑑別性，但有時因為選擇訓練資料的偏差，導致極小部分的 Candidate Tokens 的色情傾向值偏高或偏低，此時可以利用小部份的網頁來做前測 (pre-testing)，再根據前測的結果找到導致誤判的 Candidate Tokens，並將之去除。至於細部調教的程度，則視每個人對誤判的忍受程度不同而有所差異。在調校過後剩下的字詞稱為 Distinctive Tokens，儲存在 Token Database 中，做為分類步驟使用的斷詞字典。

四、分類(Classification)

這個步驟旨在測試系統在經過上述訓練後是否能有效鑑別色情與非色情網頁。給定一個測試網頁，我們首先去除 HTML 標籤，然後對文字內容做斷詞，找出出現在 Token Database 中單一個別字詞，以及每個字詞對應的色情傾向值(Porn Tendency)，由於色情傾向值是相對比率的概念，所以色情傾向值為介於 0 到 1 之間的實數值。接著，我們採用 Robinson (2003) 提出的卡方分配方法來產生色情指標值，用到的幾個方程式如下所示。

$$H = C^{-1}\left(-2*\ln\prod_{n} f(w), 2n\right) \tag{1}$$

$$S = C^{-1}\left(-2*\ln\prod_{n} (1-f(w)), 2n\right) \tag{2}$$

$$I = \frac{1+H-S}{2} \tag{3}$$

其中 $f(w)$ 為個別字詞的色情傾向值，n 為字詞數目，$\prod_{n} f(w)$ 為最大概似估計量，$-2\ln\prod_{n} f(w)$ 為近似自由度 2n 的卡方分配 (Casella & Berger 2002)，C^{-1} 為 Inverse Chi-square Function，透過 C^{-1} 算出來的值則為 p-value。方程式(1)中，採用統計推論中的 Likelihood Ratio Test (Casella & Berger 2002)，該假設檢定如圖 5 所示，H 為 n 個字詞色情傾向值利用卡方分配特性結合求出的 p-value，代表在虛無假設(null hypothesis)：認為這些 f(w)彼此獨立，而呈現非均勻分配的條件下，有多少機率顯著推論得知這個虛無假設不成立。在實際情況中，用來描述色情網頁內容的字詞是經常伴隨出現，並非彼此獨立，所以我們預期假設檢定的結果為拒絕虛無假設。

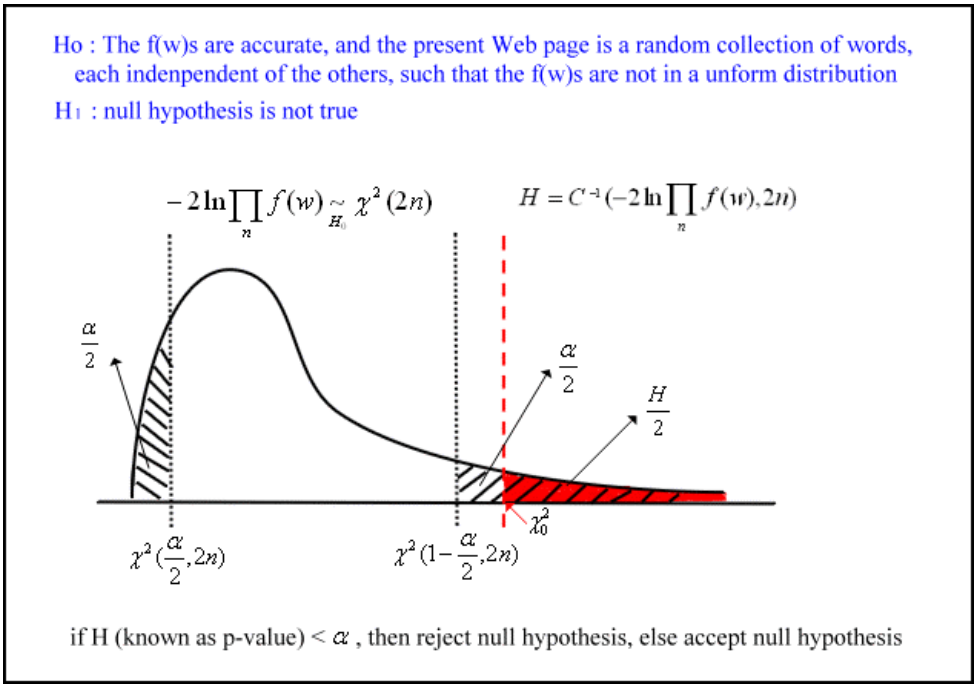

圖 5　卡方假設檢定

方程式(2)則以 *1-f(w)*(表示非色情傾向值)取代 *f(w)*，S 為 *n* 個字詞的非色情傾向值求出的 p-value。當 H、S 兩者求出之後，即可透過方程式(3)來求得色情指標值 (Indicator Value, I-value)。

方程式(3)之所以如此設計 (Meyer & Whateley 2004)，而不採用相對比率的作法(例如 $I = \frac{H}{H+S}$)，主要是因為 H 跟 S 皆為不大於 1 的實數，當某個網頁包含的色情與非色情關鍵字之個數接近時，求得的 H 與 S 值可能為 H=0.001, S=0.0001 或者 H=0.0001, S=0.001, 若以相對比率的計算方式，前者的 I 值非常接近 1，後者的 I 值非常接近 0，亦即在色情與非色情關鍵字無顯著差異時，往往會將一個“未確定”的網頁強制識別為色情或非色情網頁。

方程式(3)的實驗結果顯示，I 值會集中分佈成三個區間:當該實測網頁為明顯的色情網頁時，H 會遠大於 S，I 值在計算過後會接近 1；同樣地，若該實測網頁為明顯的非色情網頁時，H 會遠小於 S，I 值在計算後會接近 0；當實測網頁為 “未確定”(即包含的色情與非色情關鍵字之個數接近時)，I 值會落於 0.5 附近。

為了更清楚解釋方程式(1)(2)(3)的用意，底下我們分別針對“色情”、 “非色情”以及“未確定”三個種類的網頁，在 Classification 階段的數值變化用示意圖呈現。在方便闡釋的考量之下，我們將原先的雙尾假設檢定，簡化成右尾檢定，並將方程式(2)針對 *1-f(w)* 的非色情傾向值的假設檢定圖示做一個水平翻轉後，與方程式(1)並列於同一水平線上。圖 6 是針對色情網頁的色情指標值運算示意圖，方程式(1)得到的 H 值比方程

式(2)計算得到的 S 值來得大，在透過方程式(3)的計算，I 值會接近 1；同樣地，圖 7 為非色情網頁的示意圖，相對於色情網頁的數值變化，其 H 值會相對較小，S 值反而大於 H 值，經過計算後 I 值將接近於 0；比較特別的是屬於“未確定”網頁，由於內容本身沒有充份的證據顯示色情傾向的多寡，經由方程式(1)(2)會得到數值相當接近的 H 與 S 值，在計算 I 值時一加一減相互抵消的情形下，I 值會在 0.5 上下，其示意圖如圖 8。

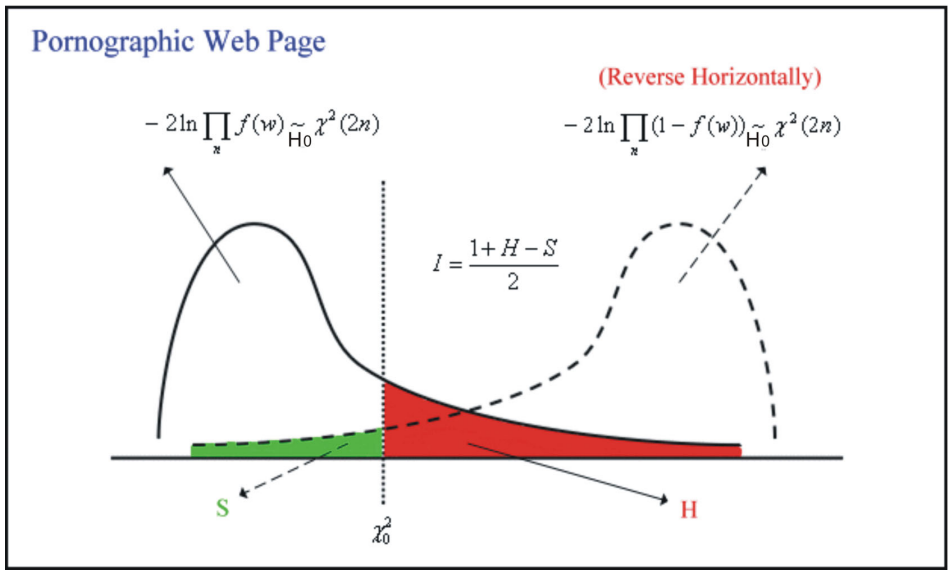

圖 6 “色情”網頁色情指標值運算示意圖

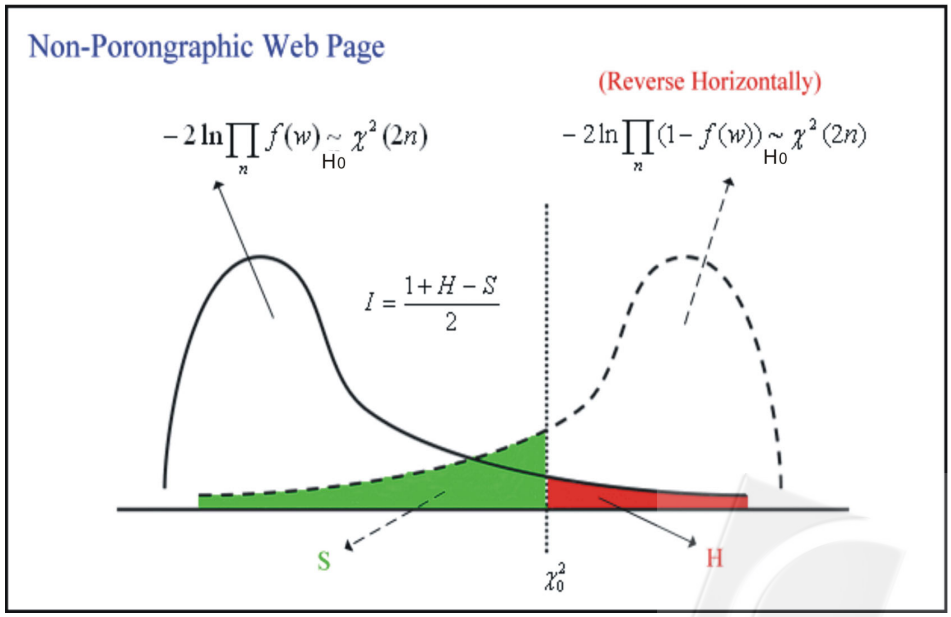

圖 7 “非色情”網頁色情指標值運算示意圖

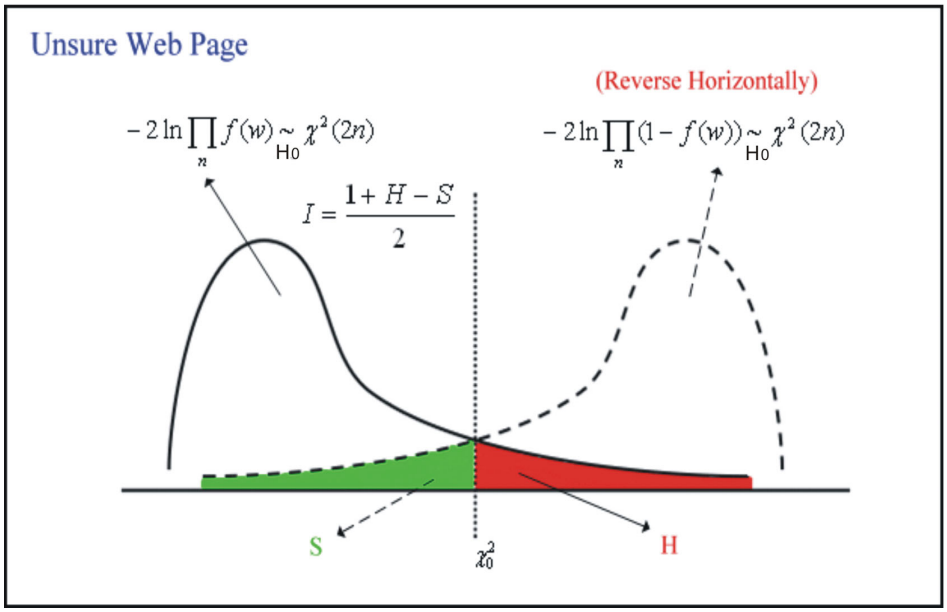

圖 8 “未確定”網頁色情指標值運算示意圖

另外，值得一提的是，並非所有斷詞出來的字詞色情傾向值都透過方程式(1)(2)(3)做運算，此處應該考慮 Effective Size Factor (ESF) (Robinson 2004； Meyer & Whateley 2004)這個概念，也就是要設定用來做計算的 *f(w)*數目的上限，例如：假若上限設為 200，當有個網頁經由斷詞之後找出 325 個 tokens，則經過由大到小的排序，本研究只取前 200 個來做方程式(1)(2)(3)的運算，求出色情指標值。考量 ESF 的原因在於，當過多的 0 到 1 之間的實數值 (*f(w) or 1-f(w)*) 連乘，將導致卡方分配的特性失效，也就是說，經由計算求出的 H 與 S 會非常的趨近於 0，即 I 值會等於 0.5，則無法判別是否為“色情”或“非色情”。經過系統模擬結果指出，在機器可識別的有效浮點數精確度，且卡方分配可以運算的條件下，對單一色情傾向值取四捨五入到小數第 4 位做計算，結果顯示字詞數目最多 500 個字詞。 若綜合考量識別精確率與時間成本，系統模擬結果顯示 150 個字詞是佳的選擇 (李龍豪等人 民 94)。

肆、系統實作與實驗評估

本節說明系統的實作情形，包含重要參數的設定與系統環境。其次，我們設計幾個實驗探討系統的精確率與正誤判率。

一、系統參數

在整個卡方色情分類的方法中，有兩個重要的系統參數，分別是門檻值組(Threshold pair)與有效字詞(Effective tokens)的個數。門檻值組 *(L, U)* 用來將 I-value

的範圍(介於 0 到 1 的實數)區隔成“色情”、“未確定”與“非色情”三個種類對應的子區間，亦即 $0 \leq I < L$ 為非色情區；$L \leq I \leq U$ 為未確定區； $U < I \leq 1$ 為色情區。另外，在給定一個測試網頁，在找到出現在該網頁的字詞以及對應的色情傾向值之後，應當結合多少數目的字詞，透過卡方分配的特性計算出色情標值，是這個以卡方為基礎的統計方法能否有效率運作的關鍵。

我們在先前的另一份研究中，提出一個系統模擬的方式，用來找到較佳的參數設定值(李龍豪等人 民 94)。該研究發現當有效字詞的數目為 150、而且門檻值組為 (0.35, 0.65)時，在平均精確率與計算時間兩相權衡之下為最具成本效益。圖 9 為在不同的門檻值與有效字詞數目的情況下，系統模擬達到的精確率分佈。

二、系統建置

本研究採用 MySQL、Apache、PHP 、GNU Wget 等軟體工具，在 Windows 2003 平台上實作系統，並針對中文與英文的網頁做內容分類實驗。系統的兩個參數則根據模擬的最佳結果做設定，分別是有效字詞為 150，門檻值組為(0.35,0.65)。本研究實作的系統也提供一個色情網頁線上分類測試機制，網址為 http://black.mis.yzu.edu.tw/antiporn/。

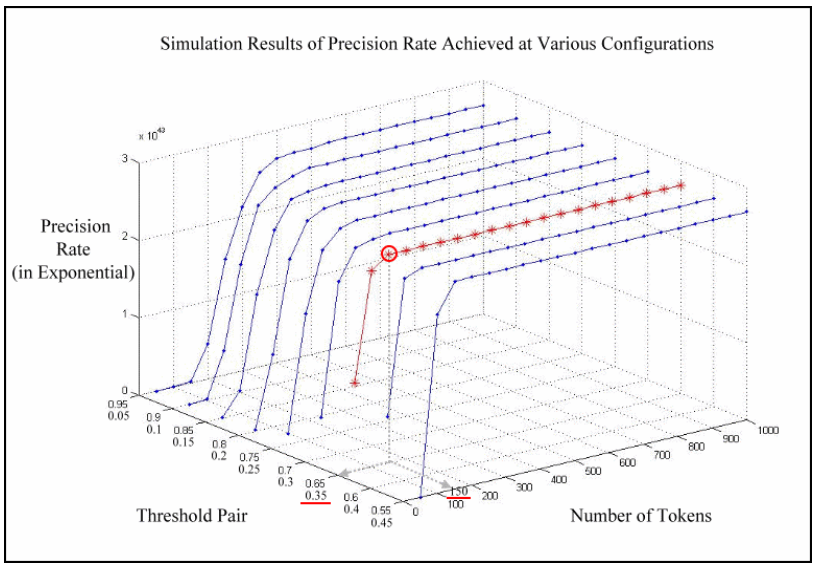

圖 9 在不同系統參數組合下的精確率 3D 立體圖

三、實驗評估

本研究以測量系統對中英文網頁分類的正誤判率 (false positive rate) 與精確率 (precision rate)來評估系統的效能。

（一）正誤判率(False Positive Rate)

本研究在 2005 年的一月到五月間，對於系統已分析的網頁進行五次隨機抽樣，每次抽樣包含中文與英文網頁各 500 筆，對於每一筆抽出的網頁，以人工判斷給予一個適當的分類。若系統判斷該網頁得到的種類與人工判斷相同，則表示系統分類正確；若系統與人工判斷的結果不相同，則視為系統判斷錯誤。

當正常網頁被系統誤判為色情時，即產生一個正誤判 (False positive)。正誤判率(False Positive Rate, FPR) 的計算如方程式(4)所示，正誤判率應當越低越好。圖 10 為 5 次隨機抽樣實驗結果的分佈趨勢，正誤判率介於 1.5% 到 2.5% 之間，而且英文網頁的平均正誤判率(1.99%) 略高於中文網頁的平均正誤判率(1.72%)。

$$FPR = \frac{Number\ \ of\ \ Non-porn\ \ Pages\ \ Misclassified}{Number\ \ of\ \ total\ \ Non-porn\ \ Pages\ \ Classified} * 100\% \tag{4}$$

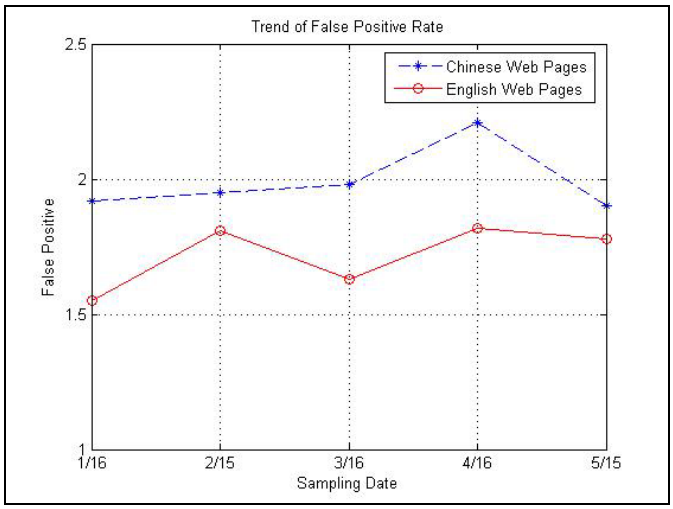

圖 10　正誤判率趨勢圖

（二）精確率 (Precision Rate)

精確率 (Precision Rate, PR) 的計算如方程式(5)所示，系統的精確率越高越好。圖 11 顯示，5 次抽樣實驗的精確率介於 95.5% 到 98.2% 之間，而且系統針對中英文網頁的平均精確率分別為 97.56% 與 96.44%。

$$PR = \frac{Number\ \ of\ \ Pages\ \ Correctly\ \ Classified}{Number\ \ of\ \ Total\ \ Pages\ \ Classified} * 100\% \tag{5}$$

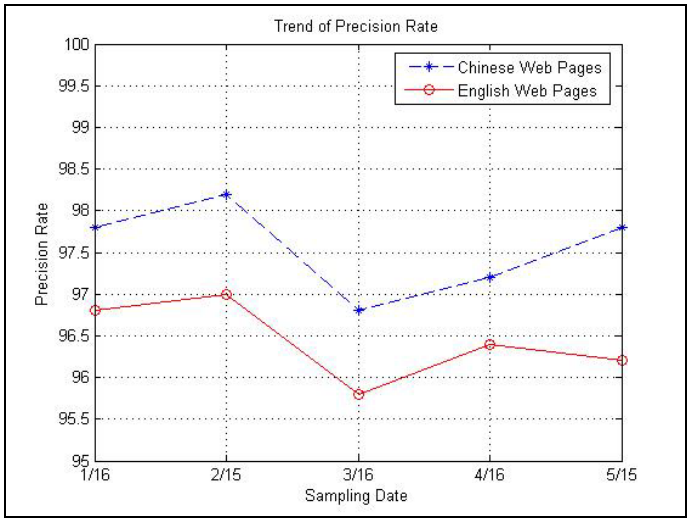

圖 11 精確率趨勢圖

這五次實驗結果中的最佳情況如表 2 所示。英文部分有 485 筆網頁正確判斷，只有 15 筆網頁判斷錯誤，英文網頁的精確率為 97%；同樣地，中文部份有 491 筆判斷正確，有 9 筆網頁判斷錯誤，中文網頁的精確率為 98.2%。中英文誤判的 24 筆網頁，大都以影像圖片的方式表現，文字部分的內容非常少，顯示我們需要採用影像分析方法來彌補卡方文字分析的不足。實驗結果與相關研究 Lee et. al. (2002& 2003) 比較，本研究的系統花費訓練時間較短，英文部分的精確率約高出 2%，而且本研究的系統還可分類中文網頁。

表 2 中英文網頁分類的精確度比較

Language	Number of URLs	Correctly Classified	Incorrectly Classified	Precision rate
English	500	485	15	97.0%
Chinese	500	491	9	98.2%

伍、結論與未來展望

本節總結本文的成果與貢獻，並探討出後續可以進行的研究方向。

一、結論

本研究針對不當資訊中的網路色情的範疇，提出一個偵測色情網頁的方法，針對網頁的文字內容，藉由卡方分配的特性結合 p-value 假設檢定，對網頁計算一個介於 0

到 1 之間的 I 值。在大量測試後，發現 I 值的分佈明顯集中於三個實數區間，可以有效區分出“色情”(Porn)、“未確定”(Unsure)、與“非色情”(Non-Porn)三個類別。

本研究針對中英文網頁實驗結果顯示，該方法具有穩定的高精確率與相當低的正誤判率。本研究實作的系統對英文網頁分類的平均精確率為 97%，比相關研究來的高，而且能夠正確判斷中文網頁。

二、未來展望

我們認為仍有幾個方向可供後續改進研究：

（一）擴充多國語系

本研究目前只處理中文與英文網頁，未來將擴大處理多國語系的網頁，例如日文、韓文、西歐語系(例如法文、德文、義大利文)與北歐語系(例如挪威、瑞典、丹麥)的網頁等，蒐集更廣泛的網路色情黑名單。

（二）處理不同領域之不當資訊

目前本系統只侷限於網路色情之領域。若能依照本架構，提出能處理其他不同領域(例如暴力、毒品與賭博等分類)的一般化方法，則可以擴大本研究之成效。

致謝

本研究由國科會專題研究計畫 NSC 93-2213-E-155-035 補助，特此致謝。

參考文獻

1. 王鐵雄、陳思翰、蔡顯明、林俊男、李新林，民 93，『從眾行為在不當資訊防制上的應用』，2004 年台灣網際網路研討會(TANET 2004)，教育部電子計算機中心主辦。
2. 李龍豪、陸承志、黃威穎，民 94，『參數調校模擬於高效率的色情網頁分類機制之應用』，2005 年台灣網際網路研討會，國立中興大學主辦。
3. 林宜隆、李璘昱、劉金和、莊育秀、許盛凱，民 92，『不當資訊防制政策與管理策略之初探』，2003 年台灣網際網路研討會 (TANET 2003)，教育部電子計算機中心主辦。
4. 邱忠俊、民 88，犯罪語言學與資料檢索應用觀念之研究—以網際網路情色文學為例，中央警察大學資訊管理研究所碩士論文。
5. 邱志傑、王明習、賴溪松，民 92，『TANet 不當資訊尋與分析』，2003 年台灣網際網路研討會(TANET 2003)，教育部電子計算機中心主辦。

6. 邱志傑、王明習、賴溪松，民 93，『不當資訊防制分析』，2004 年台灣網際網路研討會(TANET 2004)，教育部電子計算機中心主辦。
7. 邱建明、民 93，結合影像與文字辨識的網路色情過濾，國立中央大學資訊工程研究所碩士論文。
8. 郭永明、民 90，利用類神經網路決定膚色色彩空間之色情影像偵測，國立成功大學電機工程學系碩士論文。
9. 楊良吉、民 90，全球資訊網過濾軟題之研究，國立台灣大學資訊工程學研究所碩士論文。
10. Anthony (2003), SpamBayes Background Reading, available online at http://spambayes.sourceforge.net/background.html
11. Arentz, W. A., and Olstad, B., "Classifying Offensive Sites Based on Image Content," *Computer Vision and Image Understanding* (94) 2004, pp.295-310.
12. Baeza-Yates, R., and Ribeiro-Neto, B., *Modern Information Retrieval*, ACM Press, 1999.
13. Balkin, J. M., Noveck, B. S., and Roosevelt, K., "Filtering the Internet: A Best Practices Model," Information Society Project at Yale Law School, September 1999, pp. 1-38.
14. Bogofilter, available online at http://bogofilter.sourceforge.net/
15. Bosson, A, Cawley, G. C., Chan, Y., and Harvey, R., "Non-retrieval: Blocking Pornographic Images," *Proceedings of the International Conference on Image and Video Retrieval*, 2002, pp.50-60.
16. Casell, G., and Berger, R. L., *Statistical Inference* (2nd edition), Wadsworth Pub. Co., 2001.
17. Chan, Y., Harvey, R., and Smith, D., "Building Systems to Block Pornography," *Challenge of Image Retrieval*, 1999, pp.1-9.
18. Duan, L., Cui, G., Gao, W., and Zhang, H., "Adult Image Detection Method Base-On Skin Color Model and Support Vector Machine," *The fifth Asian Conference on Computer Vision (ACCV)*, 2002, pp.797-780.
19. Etzioni, O., "The World-Wide Web: Quagmire or Gold Mine?" *Communications of the ACM (39:11)* 1996, pp. 65-68.
20. Goodwin, S., and Vidgen, R., "Content, Content, Everywhere…..Time to Stop and Think? The Process of Web Content Management," *Computing and Control Engineering Journal (13:2)* 2002, pp.66-70.
21. Graham, P. (August, 2002), "A Plan for Spam," available online at http://www.paulgraham.com/spam.html .

22. Hammami, M., Chahir, Y., and Chen, L., “WebGuard: Web Based Adult Content Detection and Filtering System,” *IEEE/WIC International Conference on Web Intelligence*, WI, 2003, pp.574 – 578.
23. Jiao, F., Gao, W., Duan, L., and Cui, G., "Detecting Adult Images Using Multiple Features," *Info-tech and Info-net 2001 Proceedings (ICII)*, 2001, Vol.3, pp.378 – 383.
24. Jicheng, W., Yuan, H., Gangshen, W., and Fuyan, Z., “Web Mining : Knowledge Discovery on the Web,” *IEEE International Conference on Systems, Man, and Cybernetics*, Vol. 2 , 1999, pp.137-141.
25. Kolariand, P., and Joshi, A., “Web Mining: Research and Practice,” *IEEE Computational Science and Engineering (Web Engineering)* (6:4) 2004, pp. 49-53.
26. Kosala, R., and Blockeel, H., “Web Mining Research: A Survey,” *ACM SIGKDD Explorations* (2:1) 2000, pp.1-15.
27. Lee, P. Y., Hui, S. C., and Fong, A. C. M., “Neural Networks for Web Content Filtering,” *IEEE Intelligent Systems* (17:5) 2002, pp.48-57.
28. Lee, P. Y., Hui, S. C., and Fong, A. C. M., “A Structural and Content-Based Analysis for Web Filtering,” *Internet Research: Electronic Networking Applications and Policy* (13:1) 2003, pp. 27-37.
29. Liu, L., Chen, J., and Song, H., “The Research of Web Mining,” *Proceedings of the Fourth World Congress on Intelligent Control and Automation*, 2002, pp.2333 – 2337.
30. Meyer, T. A., and Whateley, B., “SpamBayes: Effective Open-source, Bayesian Based, Email Classification System,” *First Conference on Email and Anti-Spam (CEAS)*, 2004, pp.1-8.
31. Platform for Internet Content Selection (PICS), available online at http://www.w3c.org/PICS/ .
32. Robinson, G., “A Statistical Approach to the Spam Problem,” *Linux journal* (2003: 107) March 2003.
33. Robinson, G. (April 28, 2004), “Handling Redundancy in Email Token Probabilities, Version 0.94,” available online at http://www.garyrobinson.net/2004/04/improved_chi.html .
34. Robinson, G. (May 3, 2004), “Why Chi? Motivations for the Use of Fisher’s Inverse Chi-Square Procedure in Spam Classification, Version 0.93,” available online at http://www.garyrobinson.net/2004/05/why_chi.html .
35. Ross, S. M., *Introduction to Probability and Statistics for Engineers and Scientists* (3rd edition), Elsevier Inc., 2004.
36. Schettini, R., Brambilla, C., Cusano, C., and Ciocca, G., “On the Detection of Pornographic Digital Images,” *Proceedings of SPIE, Visual Communications and Image Processing*, 2003, pp. 2105-2113.

37. Smith, D., Harvey, R., Chen, Y., and Bangham, A., "Classifying Web Pages by Content," *IEE European Workshop on Distributed Imaging*, 1999, Vol. 99/109, pp.8/1-8/7
38. SpamBayes: Bayesian anti-spam classifier written in Python, available online at http://spambayes.sourceforge.net/index.html .
39. Srivastava, J., Desikan, P., and Kumar, V., "Web Mining : Accomplishments and Furture Directions," *Proceedings US. National Science Foundation Workshop on Next-Generation Data Mining (NGDM)*, 2002, pp.51-70.
40. The Internet Contenting Rating Association (ICRA), available online at http://www.icra.org/ .
41. Torres, L. and Vila, J., "Automatic Face Recognition for Video Indexing Application," *Pattern Recognition (35:3)* 2002, pp. 615-625.