參數調校模擬於高效率的色情網頁分類機制之應用

李龍豪[1]，陸承志[2]，黃威穎[3]
元智大學資訊管理研究所
[1]longhow.lee@gmail.com
[2]imcjluh@saturn.yzu.edu.tw
[3]s937714@mail.yzu.edu.tw

摘要

在網路蓬勃發展的今日，不當資訊的偵測與防治已經成為廣泛討論的議題。在先前的研究中，我們提出運用卡方為基礎的統計演算法於色情網頁分類，本篇文章針對所提方法的系統參數部分，提出一個系統模擬的調校機制，以期在成本效益的考量之下，達到高效率的系統效能。

關鍵字: 參數調校，系統模擬，成本效益，卡方分配，網頁分類，不當資訊過濾

1 前言

隨著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[02,04] 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等。

World Wide Web Consortium [14]提出Platform for Internet Content Selection [10]，制定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)對內容自我加註分級，或是由協力廠商(3rd party)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大約僅佔 11.0%。所以由網路頻寬提供者，例如ISP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。根據行政院新聞局所研擬的網站內容分級推廣計畫[05]，主要對照國際The Internet Content Rating Association [13] 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，再比照電視分級的作法，分為普通級、輔導級、保護級和限制級[01]，明訂出現在網站上的內容應對應何種級別，例如在語言上出現明顯與性相關字眼，將被分類為限制級。

本研究主要針對不當網路內容中最為氾濫的網路色情，提出一個參數調校模擬的機制，用於卡方為基礎的統計演算法之色情網頁分類。實驗結果顯示，我們所提出的參數調校模擬機制在成本效益考量上，可以有效率提升色情網頁偵測的精確率(precision rate)，配合自動化的網頁蒐集、過濾與分類，我們可以快速蒐集大量的色情黑名單。

2 卡方色情網頁分類

我們在先前的研究中，提出運用卡方為基礎的統計演算法於色情網頁分類[09]，藉由對網路內容中的文字部分，運用統計推論中的卡方(chi-square)分配特性，對每一個網頁計算出一個介於 0 到 1 之間的色情指標值(Indicator Value, I value)，再依臨界值(threshold)將 I 值範圍分隔成三個間斷(distinct)的區間，每個區間賦予一個類別，分別是色情(Porn)、未確定(Unsure)與非色情(Non-Porn)。

2.1 系統概觀

圖 1 是卡方色情分類的系統架構。首先，我們

國科會輔助計畫編號：NSC 93-2213-E-155-035

使用 Web Crawler 從網際網路上抓取跟色情關鍵字相關的網頁，選取部份網頁做為訓練資料(training data set)，經由 Training 產生 Token Database；其餘網頁則做為測試資料(testing data set)。我們透過一個統計演算法的卡方分配(chi-square)特性，對每個測試網頁求出一個色情指標值(Indicator Value)，以根據此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 URL Black List 中。

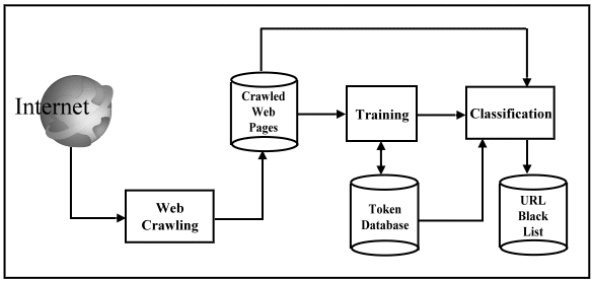

圖 1 系統架構圖

2.2 色情指標值(Indicator Value)

給定一個測試網頁，我們首先去除 HTML tags，然後依 Token Database 做斷詞，找到出現在 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 Gary Robinson 提出的 Spam Filtering 方法中的卡方分配[03,12]的特性結合 p-value 產生色情指標值，用到的幾個方程式如下所示。

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

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

$$I = \frac{1+R-G}{2} \quad (3)$$

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 [03]，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 [03]，該假設檢定如圖 8 所示，R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。類似的方程式(2)，則將f(w)取代成 1-f(w)表示非色情傾向值，G為n個個別字詞的非色情傾向值求出的p-value。當R、G兩者求出之後，即可透過方程式(3)來求得色情指標值(Indicator Value)。

2.3 系統參數

在整個卡方色情分類的研究中，有兩個重要的系統參數，分別是門檻值(Threshold Pair)和 Effective Tokens 的數目。門檻值用來決定如何將 I-value 介於 0 到 1 的實數區間，區隔成"色情"、"未確定"和"非色情"三個種類對應的子區間。另外，在給定一個測試網頁，在斷完詞找到出現在該網頁的單一個別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(Effective Tokens)透過卡方分配的特性計算出色情標值，將會是這個卡方為基礎的統計演算法能否有效率運作的關鍵。

3 參數調校模擬

我們提出一個參數調教的模擬機制，針對這兩個重要的系統參數，透過模擬的結果找到最佳的參數設定，以下將針對整個模擬的過程跟結果做詳細的介紹。

3.1 模擬環境

我們使用 MATLAB 6.5 在 Windows 環境下做系統模擬，整個模擬的環境設定如下：假設已知有"色

情”、“未確定”以及“非色情”這三個種類的網頁各3000 筆，總共 9000 筆網頁進行模擬測試；另外，假設“色情”網頁部分在經過斷詞之後得到的 Tokens 從 0.2 到 1.0 的均勻分配中隨機產生色情傾向值(f(w))；同樣地，“非色情”網頁從 0 到 0.8 的均勻分配中隨機產生 Token 的色情傾向值，而“未確定”這個部分則是從 0.2 到 0.8 的均勻分配中隨機給予色情傾向值。為了降低亂數產生器造成的偏差，我們對每一個色情傾向值(f(w))做 5 次的亂數產生，然後再取其平均當作色情情向值。

3.2 門檻值(Threshold Pair)

門檻值的設定將影響到系統的精確率(precision rate)，這裡的精確率為系統正確判斷分類的網頁數目除以已知給定分類的網頁數目。系統的門檻值(Threshold Pair)由兩個部分組成，分別是門檻下界(Lower Bound, L)和門檻上界(Upper Bound, U)，也就是說對於每一組Threshold Pair (T_i)，$T_i=(L_i, U_i)$。模擬中的L_i由 0.5 到 1.0，每次遞增 0.5；U_i則由 1.0 到 0.5，每次遞減 0.5。舉例來說 $T_5=(L_5,U_5)=(0.25,0.75)$，表 1 為模擬的結果，每個細格(cell)代表的是在該特定情況下的精確率，最後一列為在該Threshold Pair固定下，對不同的Effective Tokens數目的平均精確率。由模擬結果可知，在$T_7=(0.35,0.65)$時，整體的平均精確率為 99.95，是所以不同的門檻值設定中對系統精確率而言最好的結果。

表 1 模擬結果

Number of Effective Tokens	Threshold Pair								
	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9
50	72.41	84.25	91.01	94.71	97.03	98.36	99.14	98.88	93.05
100	88.32	94.96	97.79	98.94	99.50	99.81	99.94	99.97	99.85
150	95.20	98.49	99.41	99.76	99.89	99.97	100	100	100
200	98.22	99.51	99.83	99.95	99.98	99.99	100	100	100
250	99.41	99.87	99.97	99.99	100	100	100	100	100
300	99.76	99.94	99.99	100	100	100	100	100	100
350	99.93	99.98	100	100	100	100	100	100	100
400	99.98	100	100	100	100	100	100	100	100
450	99.99	100	100	100	100	100	100	100	100
500	99.99	100	100	100	100	100	100	100	100
550	100	100	100	100	100	100	100	100	100
600	100	100	100	100	100	100	100	100	100
650	100	100	100	100	100	100	100	100	100
700	100	100	100	100	100	100	100	100	100
750	100	100	100	100	100	100	100	100	100
800	100	100	100	100	100	100	100	100	100
850	100	100	100	100	100	100	100	100	100
900	100	100	100	100	100	100	100	100	100
950	100	100	100	100	100	100	100	100	100
1000	100	100	100	100	100	100	100	100	100
Mean	97.66	98.85	99.4	99.67	99.82	99.91	99.95	99.94	99.65

3.3 Effective Tokens 的數目

我們將表 1 中的 Effective Tokens 的數目在不同的 Threshold Pair 所得的精確率做平均，可以得到每個 Effective Tokens 的數目對平均精確率的分佈趨勢，如圖 2 所示，我們可以發現平均精確率隨著 Effective Tokens 的數目成遞增的趨勢，當 Effective Tokens 的數目超過 150 時，平均精確率已超過 99%。

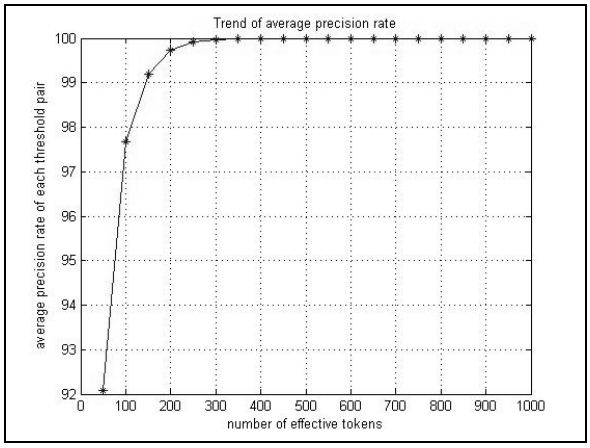

圖 2 精確率分佈趨勢圖

此外，我們也發現當較多的Effective Tokens用來計算色情指標值，將耗費較多的計算時間。為了可以有效權衡系統精確率和計算時間，我們採用以下的方程式計算Cost-Effectiveness，其中n為tokens個數，APR_n代表在所有的Threshold Pair下的平均精確率；ACT_n則是平均的計算時間。

$$Cost-Effectiveness_n = \frac{APR_n}{ACT_n} \qquad (4)$$

表 2 為 Cost-Effectiveness 的計算結果，當 Effective Tokens 的數目為 150 時，可以發現 Cost-Effectiveness 等於 7.2467，這是所有情形中最好的結果，顯然的這是模擬在無法得知何謂最佳結果(optimal solution)情形下，得到的近似最佳結果(near optimal solution)。

3.4 最佳參數設定

根據上述的模擬結果，我們可以發現當 Effective Tokens 的數目為 150 時且 Threshold Pair 為(0.35,0.65)時，在平均精確率和計算時間耗費之間，兩相權衡之下為最具成本效益的參數設定。

圖 3 為 Effective Tokens 的數目 150 時的模擬直方圖。其中"色情"部分的 3000 筆網頁，經過計算之後的色情指標值(I-value)，絕大部分近似 1(大於 0.8)；同樣地，"非色情"網頁的 3000 筆，色情指標值(I-value)則大部分靠近 0(小於 0.2)；而"未確定"的 3000 筆網頁則位於 0.5。若將 Threshold Pair 設定為(0.35,0.65)，則系統的精確率為 100%。

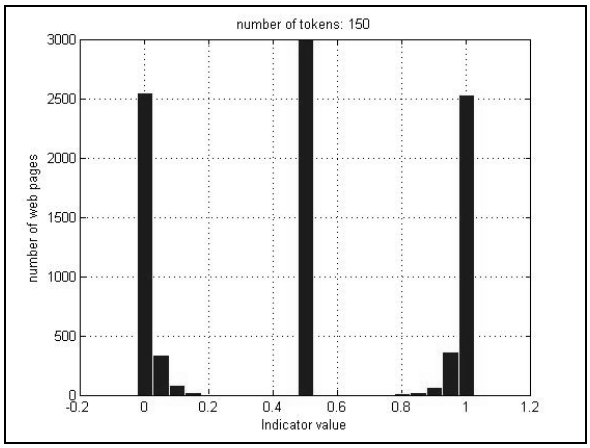

圖 3 Effective Tokens 的數目為 150 的模擬直方圖

4 系統實作與測試

我們在 Linux Fedora Core 2 環境下，採用 MySQL 4.0.20、Apache 1.3.31、 PHP 5.0.1 和 GNU Wget 1.9 [06] 將卡方色情分類的系統實作，並將重要的系統參數 Threshold Pair 設為(0.35,0.65)；Effective Tokens 的數目設定為 150，也就是說當給定一個測試網頁，若斷詞之後的單一字詞不超過 150 個時，則將其對應的色情傾向值全部用來計算色情指標值；若斷詞之後超過 150 個單一字詞，則將對應的色情傾向值由大致小排序，取前 150 個結合卡方分配計算 I-value。

我們將事先蒐集的色情關鍵字，用程式自動輸入搜尋引擎，將搜尋結果中網址的網頁抓取回來，其中部分當作訓練資料，另一部分當作測試資料。我們在測試資料中，選取中英文網頁各 500 筆，事先人工檢查給定正確分類，然後給系統作分類判斷。實驗結果如表 3 所示。

表 3 分類精確率

語言	網頁數目	判斷正確	判斷錯誤	精確率
英文	500	486	14	97.2%
中文	500	482	18	96.4%

在英文部分有 486 筆判斷正確，14 筆判斷錯誤，精確率為 97.2%；中文則有 482 筆判斷正確，18 筆判斷錯誤，精確率為 96.4%。判斷錯誤的 32 筆網頁，經過詳細檢查發現，網頁內容主要以圖片方式呈現，極小部分的文字資訊，系統無法有效判斷分類，這部份可能需要使用影像處理的技術輔助分析。跟 Lee et al. (2002 & 2003) 以類神經網路做文字內容分析的研究相比較，在精確率上提升約 2.2% (英文部分)，再者，我們也可以處理中文網頁。

表 2 Cost-Effectiveness 的計算結果

Number of Effective Tokens	Average Precision Rate of Each Threshold Pair	Average Computational Time (Seconds)	Cost Effectiveness
50	92.09	13.0656	7.0483
100	97.68	13.5624	7.2023
150	**99.19**	**13.6876**	**7.2467**
200	99.72	14.1438	7.0504
250	99.92	14.2312	7.0504
300	99.97	14.5000	6.8949
350	99.99	14.7720	6.7689
400	99.99	14.8970	6.7121
450	99.99	14.5844	6.8560
500	99.99	14.6780	6.8122
550	100	15.4312	6.4804
600	100	32.9344	3.0363
650	100	41.6688	2.3999
700	100	42.9876	2.3263
750	100	43.4344	2.3023
800	100	43.8158	2.2823
850	100	44.5750	2.2434
900	100	44.5970	2.2423
950	100	48.6126	2.0571
1000	100	57.7188	1.7325

5 結論

我們提出一個參數調校模擬的機制，用於先前提出的以卡方色情網頁分類系統，實驗結果顯示經由模擬結果得到的近似最佳參數設定值，在成本效益的考量下，可以使整個系統達到高效率，系統的精確率高達 96%以上。

未來我們計畫處理不同領域的不當資訊包含賭博、暴力、犯罪等等。另外，除了中文、英文之外，打算對其他語系做處理，包含東亞語系(日文、韓文)、西歐語系(德文、西班牙文、義大利文等等)，

企圖蒐集更完整的不當資訊黑名單，可以做為網路內容分級與資訊過濾的依據。

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. B. J. Bushman, and J. Cantor, “Media Ratings for Violence and Sex,” American Psychologist, Vol. 58, No. 2, pp. 130-141.
2. J. M. Balkin, B. S. Noveck, and K. Roosevelt, “Filtering the Internet: A Best Practices Model,” Information Society Project at Yale Law School, September 1999, pp. 1-38.
3. G.. Casella, and R. L. Berger, (2001), Statistical Inference (2nd edition), Wadsworth Pub. Co.
4. S. Goodwin, and R. Vidgen, “Content, Content, Everywhere…..Time to Stop and Think? The Process of Web Content Management,” Computing & Control Engineering Journal , Vol. 13, No. 2, 2002, pp.66-70.
5. Government Information Office, Republic of China, Project For Promoting Internet Content Rating System, available online at http://info.gio.gov.tw/public/Attachment/451214545571.doc
6. GNU Wget Project, available online at http://www.gnu.org/software/wget/wget.html
7. P. Y. Lee, S.C. Hui, and A.C.M. Fong, "Neural Networks for Web Content Filtering," IEEE Intelligent Systems, Vol. 17, No. 5, 2002, pp.48-57.
8. P. Y. Lee, S.C. Hui, and A.C.M. Fong, "A Structural and Content-Based Analysis for Web Filtering," Internet Research: Electronic Networking Applications and Policy, Vol. 13, No. 1, 2003, pp. 27-37.
9. L. H. Lee, ”A Web Content Classification for Pornographic Blacklist Generation,” Master thesis, Department of Information Management, Yuan Ze University, 2005.
10. Platform for Internet Content Selection (PICS), available online at http://www.w3c.org/PICS/ .
11. G.. Robinson, “A Statistical Approach to the Spam Problem,” Linux journal, Vol. 2003, issue 107. pp.1-9.
12. G.. Robinson (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
13. The Internet Contenting Rating Association (ICRA), available online at http://www.icra.org/
14. World Wide Web Consortium (W3C), available online at http://www.w3c.org/ .