IEBLOCKER 個人版不當網站防制系統

邱志傑 王明習 謝錫堃 賴溪松*

國立成功大學 計算機與網路中心

***國立成功大學 電機工程學系**

701 台南市大學路一號

電話(06)2757575 轉 61024 傳真(06)2368855

聯絡人:chiucj@mail.ncku.edu.tw

摘要

台灣學術網路(TANet)是以學術及教育為主之網路使用環境，在開放性的網際網路空間下，仍存在著許多的不當網站(頁)，可能造成身心發展尚未成熟之學生道德觀有所偏差。要如何防制這些不當網站在台灣學術網路中出現，避免 TANet 成為不當資訊傳播的媒介，進而產生不良之教育示範及影響，將是不可忽視的重要問題。本文除了自動搜尋可疑之不當網站及主動過濾外，也建立了 TANet 不當網站資料庫；但許多家長仍無法防制學童在家裡透過個人電腦連接不當網站，因此本研究也開發了 IEBLOCKER 個人版不當網站防制系統，並且免費提供使用者下載及使用。

關鍵詞：不當網站資料庫、個人版網站防制。

1. 前言

1.1 使用人數統計與不當網站分佈

隨著網際網路技術的進步以及網路使用人口持續的成長，上網瀏覽網站已成為最常使用的網路服務，目前網路速度愈來愈快，以往受限於檔案太大無法獲得的資料如圖片、多媒體等，現今變的隨手可得，也使得網路世界變的更豐富生動。根據 TWNIC 的 2004 年上網人數報告中指出[1]，至 2004 年 7 月中旬為止，台灣地區上網人口已達約 1,274 萬人，上網率達 56.49%。TANet 是由 12 個區域網路中心、25 個縣市教育網路中心及各級學校校園網路結合形成之整體台灣學術網路，日前統計連線單位約 4906 所，使用人數超過 346 萬人[2]。電腦硬體普級化後增加了使用者的數量，網路使用的年齡層也逐漸下降，顯示全民高度的資訊化，但這並不表示網路上的資訊都適合所有年齡層瀏覽，尤其是色情網站及賭博廣告充斥在網路世界中，這些資訊都很容易在網路上找到，可是這些資訊並不適合未成年的學童瀏覽及觀看。

在 2004 年 6 月的 Internetpolicy 統計資料中指出[3]，全球色情網頁大約有四千六百多萬的網頁，如表 1 所示。而隨著網路頻寬的增加，瀏覽網頁不再受限於以往過慢的網路頻寬，反而更以多媒體聲音及動畫影片來突顯網頁的豐富性。

表 1　Internetpolicy 分析全球色情網頁

地區	網頁數量
歐洲	28,430,600
太平洋洲	12,352,600
亞洲	3,193,000
拉丁美洲	1,048,600
非洲	389,400
加拿大	283,600
哥倫比亞	255,000
中東	77,800
總數	**46,030,600**

1.2 教育部對不當網站之防制措施

網路已經融入於現今青少年的生活之中，不論是因為學業上的需要，或者是交友傳遞訊息，網路都時時刻刻在父母和師長關注不到的地方影響著他們。如果這些含有不適合學童存取的網站不在現階段加以控制，那麼暴露在色情、暴力、賭博毒品及藥物濫用的污染下，因而影響心智和行為的青少年將越來越多，而國家又不知要付出多少社會成本才能挽救回這些國家未來的主人翁。

教育部於 2003 年 8 月邀集各專家學者共同定訂「臺灣學術網路(TANet)南區拒絕存取資訊之網站(頁)分類審議原則」[4]，並且於九十二年十月二十日起於全台各縣市區網中心採兩種防制架構，由北中南三區網中心正式執行防堵不適合存取網站，分別由台灣大學、交通大學及成功大學進行相關研究，再由教育部技術服務中心進行相關技術整合及資料庫整合，共同建立 TANet 自有之不當網站資料庫及防制架構，至今成效良好，教育部之整體防制架構如圖 1 所示。

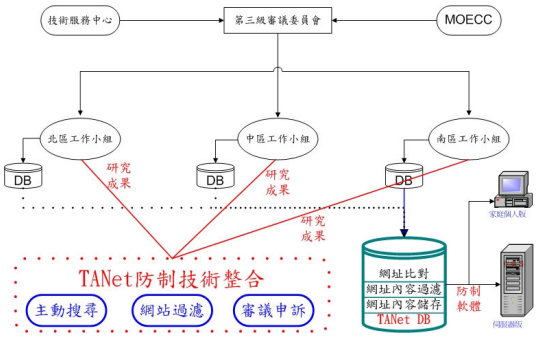

圖 1 教育部電算中心主導 TANet 技術整合

2. 不當網站資料庫系統

2.1 系統架構

傳統搜尋不當網站的方式，主要是透過搜尋與過濾兩個區塊，而本文分別以智慧型搜尋引擎代理器自動搜尋可能為不適合存取之相關網站；再透過網站分析系統(Website Analysis System)去擷取該網站本身之目錄結構，並且下載其網站完整內容，之後再利用網站分析核心(Website Analysis Core)分別計算相關資訊並且產生七大資料庫，它們分別為圖檔連結資料庫、網站名稱資料庫、內部相關網頁連結資料庫、外部相關網頁連結資料庫、關鍵字詞資料庫、中文詞彙資料庫、網站分級資料庫等七大資料庫；接著分別由關鍵字詞代理器(Keyword Agent)、圖片偵測代理器(Graphic Agent)及網站連結代理器(Link Agent)等來給予該網站合適的分類向量，將以上計算出來的向量評分結果，轉交給網站評分及分類系統(Website Rating and Classifying Engine)，透過 SVM(Support Vector Machine)演算法來分類是否為不當網站，系統架構圖如圖 2 所示。

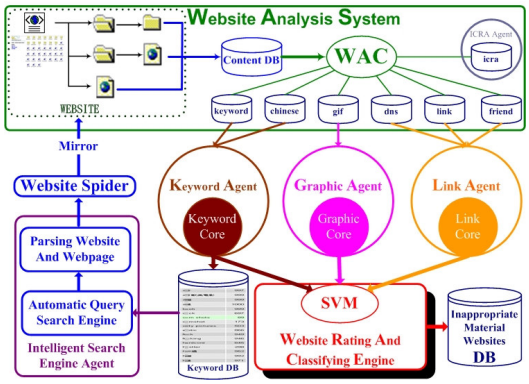

圖 2 不當網站資料庫系統架構圖

2.2 智慧型搜尋引擎代理器(Intelligent Search-Engine Agent)

本文採用目前普遍率極高之 GOOGLE 搜尋引擎來查詢可疑之不當資訊相關網頁，而智慧型搜尋引擎代理器會列出與查詢字串相關之不當關鍵字詞，並且輸出與 GOOGLE 搜尋結果同步之網站名稱及網址路徑，搜尋結果如圖 3 所示。

圖 3 智慧型搜尋引擎搜尋畫面

2.3 網站分析系統(Web Analysis System)

網站分析系統的主要功用為自動擷取該網站下所有內容，並且快速整合該網站下所有符合 htm、html、txt、asp 及 php 的網頁匯整成一個網站內容資料庫(Content Database)，並且在每個網頁切換頁加入特殊之換頁符號，提供日後判斷該網站有多少網頁之依據。如圖 4 所示，將所有特定副檔名之網頁整合成一個檔案資料庫(Content Database)之後，會立刻再透過網站分析核心(Website Analysis Core, WAC)處理整合後的資訊，產生該網站之七大資料庫。

由於網站分析系統需要大量的計算與網路下載時間，因此 WAS 可批次同時擷取約一百個網站的網頁資訊，系統本身之運算負載率維持在 40%~60%，如圖 5 左所示為 WAS 平行擷取網站內容之畫面。而 WAC 也支援平行處理，如圖 5 右所示為平行處理四個 WAC 之畫面，估計可節省約四分之三的計算時間。

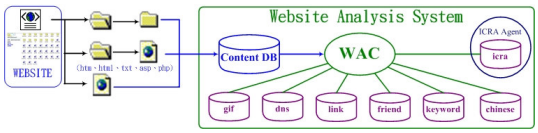

圖 4 網站分析系統架構圖

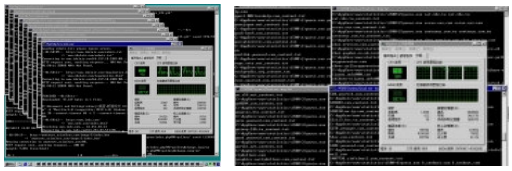

圖 5 WAS 與 WAC 支援平行處理

2.4 關鍵字詞代理器(Keyword Agent)

而關鍵字詞代理器(Keyword Agent)主要採用向量空間(Vector Space)的方法來表示每個網站，並且單獨計算每個網站的不當強度(Inappropriate intensity)及分析其中的關鍵字詞資料庫，關鍵字詞資料庫採用資料庫比對法及網頁擷取兩種方式統計。

不當關鍵字詞選取完畢後，開始計算所有關鍵字詞之權重，本文採用一般文件分類常用的 TF-IDF 演算法來計算不當關鍵字詞之比重，並且將計算出來的關鍵字詞權重儲存在資料庫中，提供後面之網站評分及分類系統(WRACE)之參考依據，計算網站關鍵字詞權重流程圖如圖 6 所示。

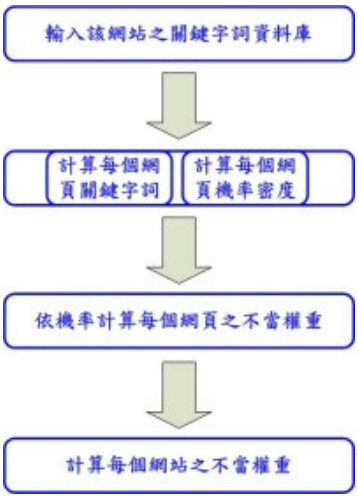

圖 6 計算網站之關鍵字詞權重流程圖

2.5 圖片偵測代理器(Graphic Agent)

圖片偵測代理器的權重計算方式為計算該網站中每一個目錄下所有圖片的膚色比例，在計算圖形偵測代理器的權重部份，利用膚色強化及動態人種膚色來減少膚色判斷之誤差，如圖 7 所示，左圖為原始影像，右圖為膚色強化及動態人種膚色處理之結果(該圖部分特徵以霧化處理)，接著計算該網站下的每個目錄的膚色比例，並且加總後為該網站的圖片代理器的權重分數，圖片偵測代理器權重計算方式流程圖如圖 8 所示。

圖 7 膚色強化及動態人種膚色處理之結果

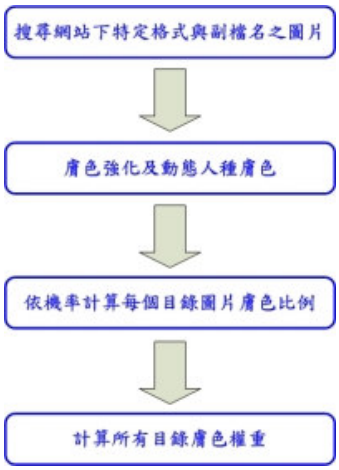

圖 8　圖片代理器權重計算流程圖

2.6 網站連結代理器(Link Agent)

市面上之搜尋引擎會利用本身的搜尋機制及評分方式來給予每個網頁一個適當的分數與順位，在[5]的研究中很仔細的針對先前搜尋引擎常用之 HITS 演算法及 PageRank 演算法的優缺點做相關比較，但是由於不當網站的分類標準比一般網站分類更為單純，許多權重微調值都可以簡化或忽略。而本文之連結代理伺服器除了分析本身網站連結到哪些網站外，也會分析該網站被哪些網站連結，有別於目前一般市面之搜尋機制都只有考慮該網站連結到哪些網頁的單向做法，如圖 9 所示，該網站有連結到一些其他的網站為 Link-Out，若該網站被一些其他的網站連結為 Link-In，分別計算每個網站被 Link-Out 及被 Link-In 的分數相加後為該網站之連結總分。

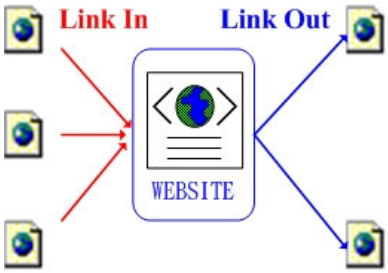

圖 9　網站連結示意圖

2.7 網站評分及分類系統(Website Rating and Classifying Engine)

每個網站分別經過關鍵字詞代理器(KA)、圖形偵測代理器(GA)及網站連結代理器(LA)後，分別計算出該網站的權重向量，此向量代表著該網站的資料屬性，這些屬性是以數字表達。在做 svm 分類之前，我們先針對訓練資料作 scaling 的動作。做 scaling 的優點有下列兩點：一是避免數字範圍大的屬性影響數字範圍小的屬性，使得分類結果不正確。二是可以避免計算上的困難。因為 svm 需作大量的向量內積的運算，太大的數值會造成計算上的困難。通常兩類的屬性值介於-1 到 1 之間(或是 0 到 1 之間)，同樣的在計算測試資料前也需要先做 scaling 的動作。

3. IEBLOCKER 個人版防制系統

3.1 動機與目的

現在是個資訊爆炸的時代，身為父母想要讓自己的小孩既能學習電腦的使用技巧而跟上時代，又要避免小孩子因為 Internet 上過於充斥的各種不設限資訊而受影響，甚至是沉迷於不適合存取網站是一件很麻煩的事情。本系統希望能夠建立一套有如〝防毒軟體〞般的機制，將不適合存取網站阻擋於個人電腦之外，藉以還給網路使用者一個乾境的網路空間以及維護青少年在心智上的正常發展。

3.2 軟硬體需求

本系統適合的作業環境為 Windows 2000/XP，其硬體方面並無特別的限制。本系統是採用 Microsoft .Net 語言開發，並且該程式會自動外掛於 Internet Explorer(以下簡稱 IE)上。根據成功大學所提供之單日使用者瀏覽器分布顯示，使用 IE 瀏覽器之民眾超過七成以上，其它廠牌瀏覽器分布如表 2 及圖 10 所示。

表 2　使用者瀏覽器一覽表

MSIE	1401644	72.3%
KKman	465210	24.0%
FireFox	36879	1.9%
MyIE2	10404	0.5%
Opera	1042	0%
Others	22045	1.1%
Total	1937224	100%

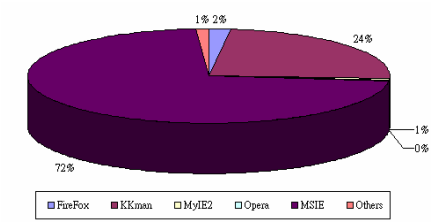

圖 10　使用者瀏覽器分布圖

3.3 IEBLOCKER 功能架構及功能簡介

當使用者透過 IE 瀏覽器上網的同時，IE 瀏覽器會將使用者預瀏覽之網頁傳送到 IEBLOCKER 機制做檢查，即時偵測使用者預瀏覽之網頁是否與內建之不當網站資料庫相符合或網頁內容含有不當網站分級標籤，若符合 IEBLOCKER 之偵測特徵，IEBLOCKER 會將使用者預瀏覽之網站、使用者 IP 及瀏覽時間回傳到紀錄伺服器(Log Server)，紀錄伺服器將使用者資訊儲存完畢之後，再分別給資料庫阻擋或分級標籤阻擋給予不同之阻擋網頁，本系統架構如圖 11 所示。

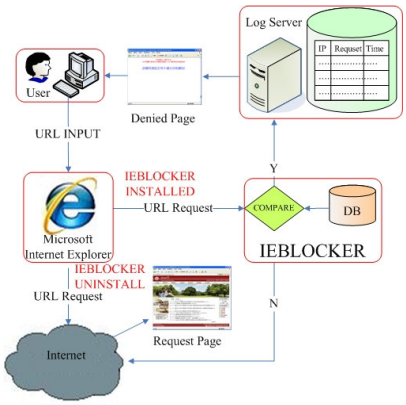

圖 11　IEBLOCKER 系統架構圖

本系統不具有使用者介面，使用者不用學習如何去使用此系統，當系統在運作時，使用者並看不到有關此系統的任何訊息，它也並沒有複雜的操作方式，相關說明及程式下載可參考 IEBLOCKER 網站[6]說明。

4. 研究成果

4.1 資料庫數量及阻擋成效

本研究在不當資訊資料庫的搜集方面，目前已經搜集超過十萬筆之不當網站資料庫，預估至少可阻擋約一百五十萬筆之不當網頁。根據中正大學瀏覽不當網站行為顯示[7]，利用前五百大不當網站瀏覽排名就可以阻絕約 95% 之不當網站，為了提供快速簡易的阻擋機制，依據現實網路存取狀態，由本系統取出排名前一千大不當網站排名，提供 IEBLOCKER 防制不當網站。

IEBLOCKER 自九五年三月發佈至九五年七月底，共有 1625 人次下載該軟體，在阻擋不當網站的次數一共有 8088 次，阻擋網站內容含有不當分級標籤共有 1632 次，一共有 9720 次阻擋使用者瀏覽不當網站之紀錄。在阻擋紀錄資料庫中共有 976 個網站被阻擋，被阻擋的前十大網站排名如表 3 所示。在阻擋紀錄資料庫中有 1572 個 IP 含有阻擋訊息，最常被阻擋的 IP 排名如表 4 所示。

表 3 被阻擋的網站排名

被阻擋的網站	次數
www.t[redacted]a.com	604
www.i[redacted]	418
photo.p[redacted]me.com.tw	369
bbs.we[redacted].com	303
www.l[redacted]com	265
www.e[redacted]ay.com	226
mypap[redacted]home.com.tw	181
www.t[redacted]nkiss.com	177
www.r[redacted]lay.com	143
ejokein[redacted]home.com.tw	138
lo-lo-l[redacted]n	124

表 4 被阻擋的 IP 排名

被阻擋的 IP	次數
220.1[redacted]1.28	290
220.1[redacted]76.230	225
220.1[redacted]76.229	218
140.1[redacted].182	216
220.1[redacted]1.31	155
220.1[redacted]1.29	140
61.62.[redacted]2	137
61.58.[redacted]133	118
61.58.[redacted]125	106
60.24[redacted]197	103
125.2[redacted]243	75

4.2 提供申訴檢舉網站

為了避免使用者認定之正常網站被 IEBLOCKER 所誤擋，本系統提供線上申訴機制[8]，提供使用者申訴被誤擋之網站，至今尚未有任何 IEBLOCKER 申訴案件。

不當網站常常透過垃圾郵件之方式散播其網站內容，因此本系統也提供使用者檢舉之管道[9]，並且定期更新 IEBLOCKER 版本，以確保其運作成效，本系統自 95 年 3 月發佈至 95 年 7 月底，共有 162 件檢舉案件，分別由 54 位使用者熱心檢舉，檢舉者之 EMAIL 排名如表 5 所示。

表 5 檢舉者的 EMAIL 排名

檢舉者的 EMAIL	次數
sure54ja[redacted]@yahoo.com.tw	34
juice@k[redacted].tw	18
sammy@[redacted]du.tw	18
zenkong[redacted]u.edu.tw	9
bsa0122[redacted]ahoo.com.tw	8
rdgeng@[redacted]idc.com.tw	7
andy562[redacted]ahoo.com.tw	4
u882132[redacted]untech.eud.tw	4
chchsiao[redacted]ail.educities.edu.tw	3
asoa600[redacted]ahoo.com.tw	3

5. 結論

本文除了建置不當資訊資料庫系統外，也開發 IEBLOCKER 個人版不當網站防制系統，可以保護使用者透過個人電腦瀏覽不當網站，並且提供使用者免費下載。此外，也提供申告網站誤擋及不當網站檢舉機制，避免有誤判之行為或遺漏之不當網站。

而保護學童避免瀏覽不當網站的最好的方式是「陪小孩上網」，讓學童了解父母的關愛，並且將電腦放置在客廳等家中「交通要衝」，不要讓小孩躲在房間中上網，同時教育小孩關於網路安全的常識，只有學童家長的關心才能真正保護學童的身心健全發展。

6. 誌謝

感謝教育部電算中心經費補助。

7. 參考資料

[1] http://mag.udn.com/mag/dc/storypage.jsp?f_ART_ID=5254

[2] 劉金和, “台灣學術網路之現況與展望”, 2001. 11

[3] http://gipi.typepad.com/internetpolicy/2004/06/

[4] http://web110.ncku.edu.tw/

[5] Sugiyama, K., "Refinement of TF-IDF schemes for web pages using their hyperlinked neighboring pages," ACM Hypertext, pp. 198-207, 2003.

[6] http://tanet110.ncku.edu.tw/ieblocker/

[7] 王鐵雄、陳思翰、蔡顯明、林俊男、李新林，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，台東。

[8] http://tanet110.ncku.edu.tw/mistake.php

[9] http://web110.ncku.edu.tw/report.php