以網路處理器技術建構 Gigabit 不當資訊過濾系統

陳鴻彬[*], 劉榮太[+], 陳闕民[*], 詹承勳[*], 黃依濤[*], 鄔培麟[++], 葉肩宇[*], 黃能富[*,+]
[*]威播科技股份有限公司　[+]國立清華大學資訊工程學系
[++]國立清華大學通訊工程研究所
{hbc@broadweb.com.tw, nfhuang@cs.nthu.edu.tw}

摘要

隨著網路技術的迅速發展,網際網路 (Internet) 已成為大眾攫取資訊的主要管道之一。加上寬頻網路的建設腳步加快，更促進了 Internet 上的資訊越來越豐富。然而有許多網站所提供是有危身心健康之不當資訊，尤以色情網站為數最多。如何過濾不當資訊網站，使網路資訊更為純淨，已成為網路管理 (尤其是 TANet) 上的重要議題。本文針對如何在 Gigabit Ethernet 網路環境裡，提供一個超高速效能的 URL 過濾技術, 過濾色情,暴力,賭博等等不當資訊，以解決當前網路上不當資訊氾濫的問題。本文所設計的過濾系統以最新進的網路處理器為基礎,可提供極為優異之 URL 過濾效能。

關鍵詞：網路處理器，URL 過濾引擎，網站自動判別引擎，不當資訊防治。

1. 前言

由於資訊網絡建設一日千里，各項硬體建設突飛猛進，使得民眾經由資訊網路取得資訊也日益便利與普遍。民眾可經由各種方式上網以攫取資訊。例如：撥接，ADSL，纜線數據機 (Cable Modem)，寬頻社區，校園網路，政府網路，企業網路，網咖等等。然而，也使得許多不當的資訊更容易被民眾所攫取。尤其是未成年的青少年，更可輕易的經由校園網路，或是網咖上網，接觸到色情網站等,容易造成許多社會問題。國內 TANet 經過多年來的發展，連線的學校已由大專院校普及到所有高國中小，而 TANet 骨幹更已升級成超高速乙太網路 (Gigabit Ethernet)。而在 TANet 建設的成就之下，所伴隨的問題便是各級學校的學生更方便藉由校園網路瀏覽一些色情、暴力、毒品、犯罪或賭博性網站。而對於大多數的青少年學生而言，正值身心發展尚未健全階段，許多事情均無法獨立判斷，建立正確觀念。因此，很容易受到不當資訊的影響，妨礙身心健全發展。而青少年又是國家未來的棟樑，長此以往，對國家社會勢必造成不良影響，嚴重的話，將使現今國家發展的成果受到阻礙。因而，如何防範不當資訊的散播，尤其是在 TANet 上的擴散，已成為國內網路發展及網路管理的一項重要課題。而不當資訊的防治，以色情網站的過濾為首要。底下將先介紹一些傳統防範不當資訊網站傳播的過濾機制，然後再介紹可在 Gigabit 骨幹網路上，具 Gigabit 處理速度效能的 URL 過濾技術，以便過濾不當資訊網站。此外，本文更將介紹如何以更有效的機制，將色情網站網址收集起來，加以過濾。

2. 一般不當資訊網站過濾的解決方案

在過濾不當資訊網站的工作上，一般所採用的方式大都為軟體解決方案，依解決方式大致上可分成三種方案：瀏覽器過濾軟體、防火牆(Firewall)式過濾軟體及代理/快取伺服器(Proxy/Cache Server)過濾軟體。這三種方案都會內建不當資訊資料庫。

2.1 瀏覽器過濾軟體

採用瀏覽器過濾軟體的解決方案，是最簡易方便的方法。其做法是在 PC 工作站上安裝過濾軟體，或安裝具有過濾 URL 網頁地址功能的瀏覽器軟體。當使用者瀏覽網頁時，過濾軟體會先將該使用者所欲瀏覽的網址，拿來搜尋內建資料庫，若搜尋到，則拒絕該次瀏覽；若搜尋不到，則允許該筆瀏覽將網頁下載回來。至於，內建資料庫通常可分為個人版或網路版的方式。個人版的內建資料庫，安裝於使用過濾軟體的 PC 工作站內；而使用網路版的資料庫，則安裝於網路伺服器中，等待過濾引擎前來查詢。通常網路版的資料庫，會有使用者人數限制，用戶必須依照 PC 工作站的數量來購買適當的使用者人數之使用權。

採用這種過濾方式的解決方案，必須針對不同的作業系統及瀏覽器購買/安裝不同版本的過濾引擎軟體。不過，現今大部分的 PC 工作站，一般都採用 Windows 系列作業系統及 IE 瀏覽器，因此，在安裝版本方面的問題不大。然而，由於每台 PC 工作站都要安裝維護該過濾引擎軟體，此點對於擁有較多 PC 工作站的單位，在維護上較為不方便。採用瀏覽器過濾軟體方案的另一項缺點是，因軟體安裝在 PC 工作站上，通常會遭使用者有意無意的將該軟體移除或解除功能，因而不能有效的過濾。

一般而言，採用瀏覽器過濾軟體過濾不當資訊的方案，較適用家庭的個人用戶。通常是家長用來避免未成年子女，上網接觸不當資訊最簡便的方式。而對於中大型網路，如校園網路而言則較不適用。因此，在 PC 工作站上加裝瀏覽器過濾軟體來

過濾不當資訊網站，較常見於家庭用戶；學校用戶一般則不多見。

2.2 防火牆式的過濾軟體

防火牆式的過濾軟體，主要是將不當資訊過濾的功能加在防火牆上，或以硬體安裝過濾軟體，架設在網路出口的路徑上，進行 URL 網址的過濾檢查。內部用戶所有的網站瀏覽要求都將經過防火牆式過濾軟體的檢驗，一旦發覺是瀏覽內建資料庫中的不當資訊網站，則將此筆瀏覽過濾掉。

採用此方式的優點是可以不改變原先網路架構，且不須額外設定，所有用戶的每一個瀏覽均會被檢視。缺點則是由於每一筆網站瀏覽的要求均要比對系統內龐大的資料庫，因而系統的處理能力將受到嚴苛的挑戰。由於是軟體的處理，因而大量消耗 CPU 處理時間，將連帶影響到其他非網站瀏覽的網路連線效能，容易形成出口流量的頻頸，造成網路阻塞。更嚴重的問題是，如果網站瀏覽流量過大，極易造成系統無法負荷而當機，導致網路斷線。因此，目前採用這種防火牆式的過濾軟體來過濾不當資訊的比率較少。

2.3 代理/快取伺服器過濾軟體

目前最常用的不當資訊過濾方案，乃是採用代理/快取伺服器加上 URL 過濾軟體的解決方案。在做法上，一般採用路由器的 WCCP 流量重導(WCCP Traffic Redirect)技術，或 L4 交換器的流量重導技術，將所有對外的網站瀏覽要求導引到代理/快取伺服器，由代理/快取伺服器代為攫取網頁回來，這便是所謂的透通代理(Transparent Proxy)。在代理/快取伺服器內會內建不當資訊網站 URL 網址資料庫，有時稱為黑名單(Block List)。當代理/快取伺服器接受來自 PC 工作站的網頁瀏覽要求時，會先透過 URL 過濾軟體進行比對檢查，看看所要求瀏覽的網頁是否在黑名單之列。若在黑名單之列，則拒絕該筆瀏覽，並回傳拒絕訊息的網頁；若不在黑名單之列，則代理/快取伺服器會為其攫取所欲瀏覽的網頁，並留存一份在內部硬碟之中，以便下次其他 PC 工作站的存取。

由於許多單位都已擁有代理/快取伺服器的設備，同時也採用透通代理的技術。因此只要再把 URL 過濾軟體安裝在代理/快取伺服器之上，便可很快地作為不當資訊過濾的過濾器。因此目前採用這種不當資訊過濾方案的單位數量最多。

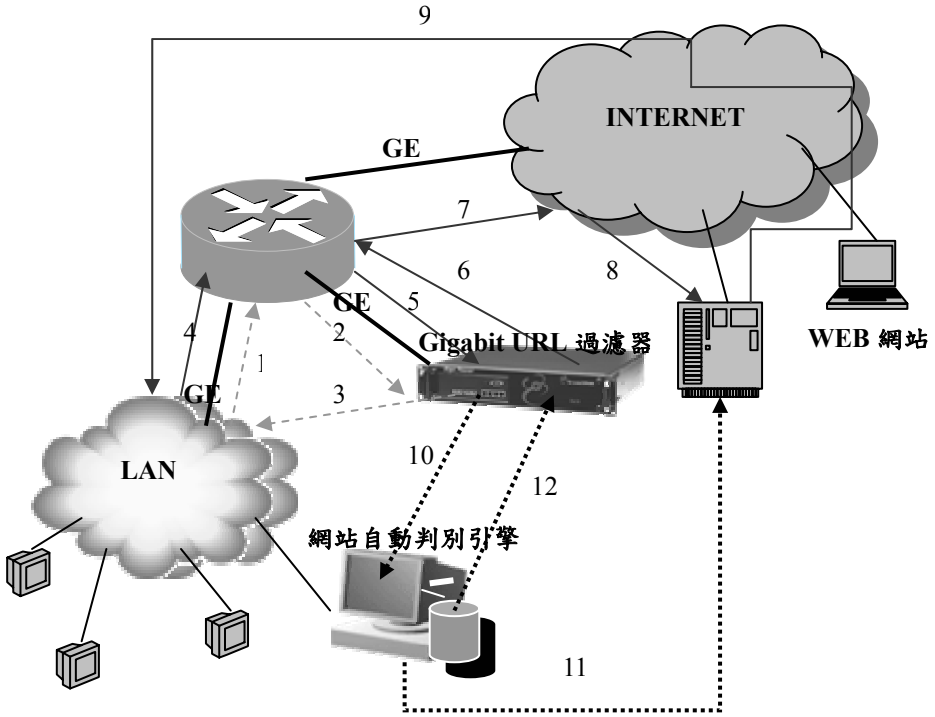
此種不當資訊過濾方式最大的問題乃是效能問題。由於原本代理/快取伺服器主要是設計來作為網頁的代理查詢及快取功能，因而在效能上並未將 URL 過濾檢查列入考量。再加上是軟體解決方案，因而每筆 PC 工作站的網頁瀏覽需求都將需比對不當資訊資料庫，因而，要消耗 CPU 的計算時間。因此，將嚴重對代理/快取伺服器的效能。尤其是一般所採用的不當資訊資料庫，通常都在數十萬筆以上。因此整體效能是否足以支援對外較高速之頻寬，將是一項嚴苛的考驗。

3. Gigabit 網路 URL 網址過濾系統

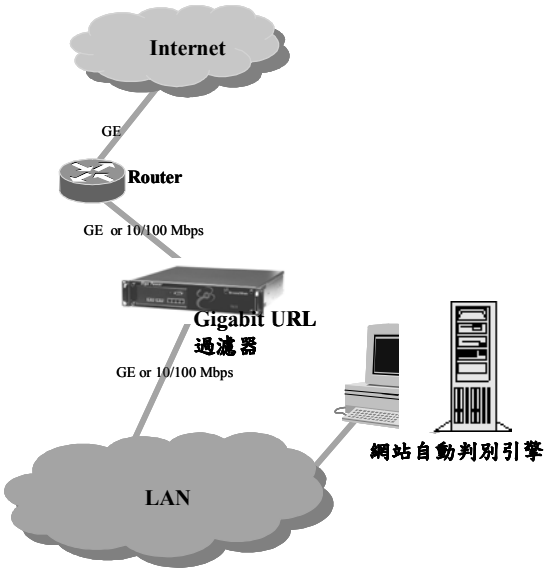
前述三種不當資訊過濾方法有其共通性，此共通性便是均為軟體解決方案，而且均內建有數萬至數十萬筆以上之不當資訊網站資料庫。因此在比對上將大量需要 CPU 的計算。其效能將受限於硬體配備及軟體技術的優劣。尤其現今網站瀏覽已佔據網路上主要交通流量，每秒所要處理網站瀏覽要求急遽增加，這些軟體解決方案在面對這樣的環境，處理上將面臨極大的挑戰，甚至可能無法負荷。尤其 TANet 骨幹已全面提升為 Gigbit Ethernet 網路，各校也積極更換校園網路成為 Gigabit 校園網路，這顯示 Gigabit 網路時代已經來臨。在如此高速的網路頻寬下，過去習慣採用的不當資訊過濾軟體已無法在這種超高速網路環境下圓滿達成任務。因此，必須有更好更新的不當資訊過濾技術及平台，才能在 Gigabit 網路下提供不當資訊過濾的服務。底下將介紹一種新的網路處理器 (Network Processor，NP) 平台技術，並在這技術上發展 Gigabit 網路不當資訊過濾技術及整體解決方案。此外，對於不當資訊網站，尤其是色情網站網址資料庫的收集，本文亦提供一套特殊解決方案，以達到有效的不當資訊過濾效果。

3.1 Gigabit URL 過濾系統架構

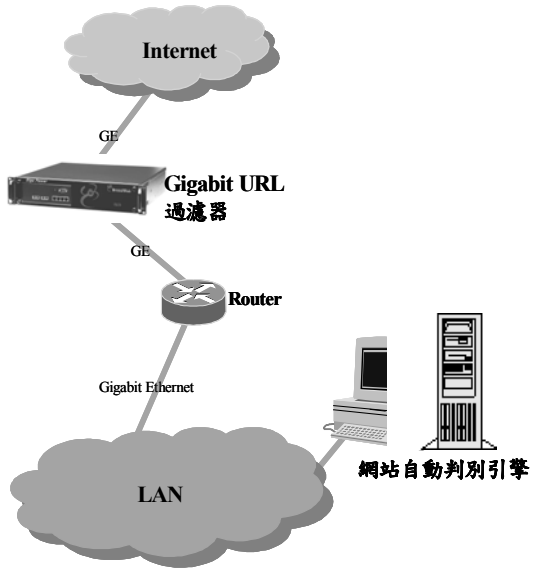
為解決軟體過濾系統在效能及速度上的瓶頸，本文將介紹可完全支援 Gigabit 網路環境的 URL 過濾系統；除此之外，本系統亦將提供一套特殊技術，以提高不當 URL 的過濾效果。本系統之完整架構，如圖一所示。本系統包含兩個主要部份，其中之一為 Gigabit URL 過濾器；另一部份則為網站自動判別引擎。Gigabit URL 過濾器可採用重導過濾模式(Redirect Mode)安裝於圖一所顯示的位置。藉由 WCCP 協定的運作，路由器會將所有的網站瀏覽封包，導引至 Gigabit URL 過濾器檢查及過濾。而網站自動判別引擎可安裝在圖一的 LAN 裡面，其功用為自動上網攫取網頁，並判斷網站是否為含有色情的不當資訊網站，並將所收集到的色情不當資訊網站列表，挹注給 Gigabit URL 過濾器，以增強整體系統的過濾效果。另外，Gigabit URL 過濾器亦可採用透通過濾模式(Pass Through Mode)安裝在 LAN 與對外路由器之間(圖二)或路由器與 Internet 之間(圖三)，檢視所有網路上的封包要求，以過濾欲瀏覽不當資訊網站的封包。此兩種架構，亦可搭配網站自動判別引擎，以增加過濾效果。



圖一. Gigabit URL 過濾系統架構(一)



圖二. Gigabit URL 過濾系統架構(二)



圖三. Gigabit URL 過濾系統架構(三)

3.2 Gigabit URL 過濾器

本系統所採用的 URL 過濾器需具備支援 Gigabit 網路的效能，因此絕非過去軟體 URL 過濾系統所能勝任。在技術上,需採用特別的平台技術。此 Gigabit URL 過濾器包括：具 Gigabit 效能的網路處理器(Network Processor)平台、微程式碼(Micro Code) 編寫的 URL 過濾引擎核心,及內建的不當資訊 URL 資料庫。

網路處理器技術是一種介於 ASIC 晶片與軟體技術之間的高階網路處理技術。其主要原理是採用專門處理網路封包的網路處理器，搭配流程控制用的 RISC CPU 與其他重要的元件形成一完整的網路處理器硬體平台。網路處理器在運作上，需撰寫微程式，並將微程式載入網路處理器直接執行。由於微程式技術，是一種專門為網路處理器所設計的類組合語言程式技術，接近於硬體技術，因此在效能上將遠快於軟體技術。在封包處理速度上，一般網

路處理器每秒可處理數百萬個封包，並達到 Gigabit 網路處理效能。以 Vitesse IQ2000 網路處理器為例，其處理能力為每秒最多可處理到 300 萬個網路封包，且最高可支援到 OC-48(2.4Gbps)網路頻寬[1]。採用網路處理器技術具有 ASIC 高速效能之優點，而微程式技術，更使得產品功能可彈性調整增加。因此，網路處理器平台，兼具 ASIC 硬體速度上的優點，及軟體功能上可更新調整的優點。

在 Gigabit URL 過濾器中，除了網路處理器硬體平台及由微程式碼所編寫的 URL 過濾引擎外，內建的不當資訊 URL 網址資料庫更是 Gigabit URL 過濾器能否發揮功效的重要元件。對於 Gigabit URL 過濾系統而言，除了要追求 Gigabit 高效能目標外，高準確性的辨識功能亦非常重要。因為唯有如此，Gigabit URL 過濾系統才可完全發揮其功效，不會成為網路障礙而讓用戶抱怨連連。因此，在 Gigabit URL 過濾系統中，內建的 URL 網址資料庫之準確性及維護更新機制與技術將影響到整體系統的功效。

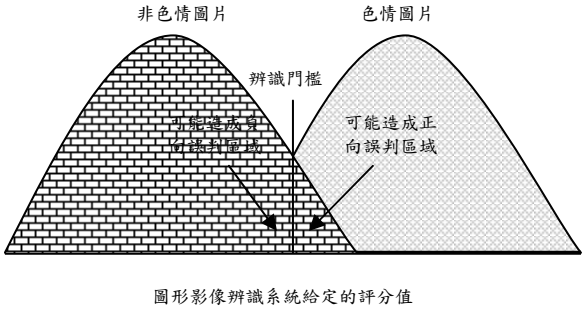
3.3 不當資訊資料庫之更新維護

對不當資訊過濾系統而言，網頁的辨別分類技術優劣，及資料庫更新方案，將影響系統的準確性，尤其是對於多變的色情網站(Porn Sites)影響更大。根據 NRC (National Research Counsel)統計，全球每天大約有 300 到 400 個新的色情網站產生，當然也有相當多的色情網站消失。因此，內建資料庫的網頁判別技術及更新維護機制，將影響 Gigabit URL 過濾系統之準確性。

3.3.1 不當資訊網頁辨識技術

一般辨別不當資訊網頁所採用的技術大都為文字辨識技術，其中最簡單的方法便是採用關鍵詞(Keyword) 比對技術。有些會加上斷詞斷句，前後文語意分析，及字根字首處理等方式以增加其準確度。此外，關鍵詞比對技術，除了比對網頁內容外，有些還運用在比對網站的網址上。採用文字辨識方式技術，最大的缺點為，無法全球化。此乃因為不當資訊網站，尤其是色情網站，各個國家地區都有，各種語言的網站都有，而每種語言都有其特殊性。因此想要完成每種語言網站的文字辨識，實在一件非常龐大的工程，亦非短期可達成。另一項文字辨識技術的缺點為許多文字被內嵌在含有圖片的影像檔案之中。因此文字判別系統無法去讀取該文字。大部份採用文字判別技術的不當資訊資料庫，其準確性並不是很高，因此大多會輔以人工判讀，以增加其準確性。

較新且有效(尤其對色情網站而言)的辨識技術，則採用圖形影像的辨別技術，來進行網站網頁的判讀。這種方式，是採用圖形影像的色澤、材質、形狀、紋理、分佈位置、對比等等參數來進行判別，在準確度通常較文字比對技術要好，尤其對於色情網頁的辨識上效果更佳。採用圖形影像判別技術，通常會將形影像以某種方法給予評分，再根據評分後圖形影像分佈在不同類別的統計分析機率，取出適當的辨識門檻(Threshold)作為分類的依據。採用此技術辨識不當網頁可不受限於網站語言種類。然而，不論影像圖形辨識技術如何進步，仍會造成如圖四所示的判別灰色地帶，而這灰色地帶則是形成正向誤判(False Positive)及負向誤判(False Negative)的誤判主要地帶。至於要如何微調辨識門檻，以增加判別的準確性及減低誤判，則是相當適合研究的課題。



圖四. 圖形影像分類識別

為提高圖形影像辨識的準確性，降低誤判，美國 VIMA 公司曾提出多形態分類演算法(Multimodal Classification Algorithm，MCA) 及概念位移演算法(Concept-Shift Algorithm，CSA) [2-3]來增加圖形影像辨別的準確性。其方法是採用訓練模型先對辨識系統作正面訓練(Positive Training)及負面訓練(Negative Training)，如圖五所示。在判別時，依照色澤、材質、對比、形狀、分佈位置等等，將每一圖形影像提煉出 150 種特性加以比對，依此可計算出特性的抽象距離(Conceptual Distance)，並給予適當的評分。

圖五 圖形影像識別系統訓練圖例

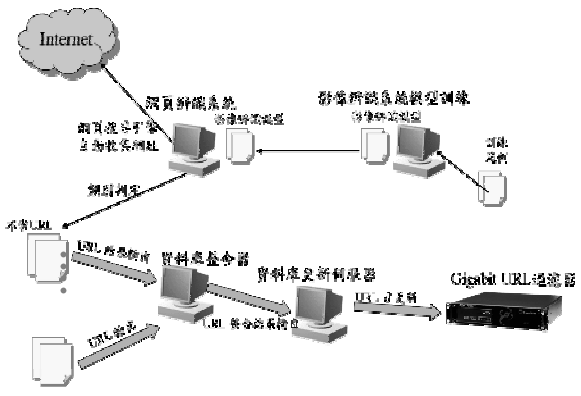
MCA 及 CSA 演算法會使用文字比對技術，來加強判斷網頁是否偏向於色情類別或偏向於非色情類別。這種輔助的判斷有助於微調辨識門檻應往色情類別提高，或往非色情類別提高。舉例而言，若文字辨識判斷出，帶有色情字眼的文字佔相當比

率，則圖四的辨識門檻將往左移動，如此，將導致圖形影像被判定為色情的機會大大提升，反之亦然。在 VIMA 所提出的實驗數據中，經過測試 5,500 張成人圖片及 3,000 張一般非色情圖片的結果顯示，這個技術在辨識門檻介於 50%的中間值時，成人圖片被準確的判斷成色情圖片的成功率達到 91.34%；而非色情圖片被準確的歸類到非色情類別的成功率亦高達 95.98%。而另一項大型的測試結果顯示，在測試 2,537,105 張色情圖片後，其準確亦高達 92.23%。這在目前各種網頁辨識技術上算是相當不錯的成果。

3.3.2 不當資訊資料庫更新維護機制

在不當資訊過濾系統中，內建不當網頁 URL 網址資料庫的持續更新維護，是確保系統持續有效運作的最重要機制。底下將提出兩種資料庫的更新維護機制，並混合使用，以確保系統持續可靠的準確性。

資料庫的更新機制可採用集中式自動更新及本地自動更新機制。圖六顯示集中式自動更新機制之流程，其中網頁辨識系統將不斷的自網路上擷取全球網站的網頁，並進行判讀是否含有不當資訊，之後定期產生新的不當資訊 URL 資料庫輸出。為使不當資訊 URL 資料庫的來源可彈性而廣泛的收取，因此在設計上多了一層資料庫整合器(Database Integrator，DI)。如此可使資料庫的來源，不受限於單一來源，以增加資料庫收集速度及功效。最後經過整合的資料庫，將透過資料更新伺服器(Database Upgrade Server，DUS)，定期自動的更新所有 Gigabit URL 過濾器內的資料庫。由於不當資訊 URL 資料庫動輒數十萬筆以上，龐大的資料量傳輸將造成系統及網路負擔。本文所提出之機制採用只更新變更的部分。包含版本之間新增之網站以及已消失的網站。這特性對於經常改變網址的色情網站資料庫而言，是非常重要的。



圖六. 不當資訊資料庫集中式自動更新機制

另一種更新機制則是在 Gigabit URL 過濾器安裝的所在地，安裝一套網站自動判別引擎，來進行本地自動更新,如圖一所示。前曾提及，每天大約會有 300 至 400 個新的色情網站產生。因此在集中式統一自動更新系統中有可能未能及時找尋出新的色情網站。甚至有些色情網站，屬於網友自行架設僅分享少數群眾的未公開網站，因而不易被發覺過濾。而採用這種本地網站自動判別引擎之更新機制，則可經由用戶之實際瀏覽行為，加以追蹤判讀其曾經瀏覽的網站是否含有色情不當資訊。如此將可確保只要用戶曾瀏覽過的網站，雖可能在第一次瀏覽時未能及時過濾，但在經過網頁自動判別引擎的追蹤後，便會被加入過濾器之資料庫，以防止其他使用者繼續瀏覽該網站。圖一中也闡述了整個過濾及追蹤的過程：

1. 使用者發出 http 網頁瀏覽要求
2. 網頁瀏覽要求被重導至 Gigabit URL 過濾器進行檢查
3. 若此瀏覽要求存在於 Gigabit URL 過濾器內建的不當資訊資料庫中，則拒絕該筆瀏覽，並回覆一個回絕網頁通知
4. 使用者重新發出 http 網頁瀏覽要求
5. 網頁瀏覽要求被重導至 Gigabit URL 過濾器進行檢查
6. 若此瀏覽要求不存在於過濾器內建的不當資訊資料庫，則將該要求轉送出去(但不代表瀏覽的網頁是正當的資訊)
7. 被轉送出的 http 網頁瀏覽要求經由路由器抵達 Internet
8. 此網頁瀏覽要求經由 Internet 抵達欲瀏覽的 Web 網站
9. Web 網站經由 Internet 送回網頁到使用者的瀏覽器，完成瀏覽過程
10. 在執行 6.的同時，Gigabit URL 過濾器複製此次瀏覽的 URL 網址，並送交網站自動判別引擎進行後續追蹤檢查
11. 網站自動判別引擎會根據所收集到的 URL，自動前往目的 Web 網站擷取網頁進行自動判讀
12. 當網站自動判別引擎檢查後發覺為不當網頁，則將該筆 URL 加入 Gigabit URL 過濾器的內建資料庫中，完成本地自動更新資料庫工作

在此過程中，網頁自動判別引擎亦可將所收集到的色情 URL 網址資料，回傳給圖六中的資料庫整合器，再集中自動更新給所有的 Gigabit URL 過濾器。如此，在整個資料庫更新個系統上，便宛如有一個分散式網頁自動收集判別更新系統，將可大大提升不當 URL 收集的能力。本系統之所以能採用這種本地更新方式，主要在於是將前述 3.3.1 中所提的不當資訊網頁辨識技術運用到網頁自動判別引擎中，如此才可在沒有人工處理的情形下，對色情網頁的判別，達到某種程度以上的準確率。□

4. 未來發展

在本文中提到了兩項重要的技術，一為網路處理器平台及微程式碼技術；另一則為網頁自動判別技術。在網路處理器平台及微程式碼技術的應用上，由於寬頻建設的發展迅速，Gigabit 骨幹網路已逐漸普及，甚至在不久的將來會演變而為 Gigabit 都會網，及企業網。因而現今許多網路設備將會逐一升級為具 Gigabit 效能之設備。而隨著 Internet 技術越發達，網路安全設備也越重要。因而，發展 Gigabit 等級的網路安全設備，如防火牆、入侵偵測防禦系統 (IDS)[4]以及 VPN 等等設備，將是趨勢所在。而這些設備，在特性上均為處理封包的設備，加上功能上均較複雜極多變性，因而非常值得採用網路處理器平台來開發。至於在網頁自動判別技術上，由於至今仍未有一套方法可完全準確的判別，因而仍有改善空間。研究更好的圖形影像辨識技術或文字語意判別技術應該值得繼續投入。或者，加上採用人工智慧及類神經技術可改善判別的準確性，這些都有賴專家學者們貢獻進一步的智慧。

5. 結語

由於網際網路的蓬勃發展，帶動了資訊傳播的革命，各式各樣的資訊在網路上唾手可得，而寬頻技術的進步更加快了此一趨勢。然而並非所有的發展均朝正面有意義的方向進行，網路上的資訊內容即是一例。現今網路上到處充斥著不良的資訊，此種資訊正負面的影響著我們下一代的身心發展，因而身為網路的一份子，有義務盡一份心力發展有用的網路資訊過濾技術，以過濾不當的網站資訊。本文便是基於此一理念而提出的一套在 Gigabit 網路上過濾不當網站資訊的完整解決方案。目的是希望藉由此技術的推行，能有助於淨化國內的網路，尤其是教育網路，以提供一個乾淨的網路學習環境，使國內網路發展朝更正面積極的應用邁進。

6. 參考文獻

[1] "IQ2000™ Family of Network Processors", *Design Manual, Revision A3.20, VITESSE Semiconductor Corporation,* 2001.

[2] Y.-L. Wu, E.Y. Chang, K.-T. Cheng, C.-W. Chang, C.-C. Hsu, W.-C. Lai, and C.-T. Wu "MORF: A distributed multi modal information filtering system (extended version).", *Technical Report, VIMA Technologies,* June 2002.

[3] Simon Tong and Edward Chang. "Support vector machine active learning for image retrieval.", *Proceedings of ACM International Conference on Multimedia.*, Pages 107-108, October 2001.

[4] 陳鴻彬 "NetKeeper 防駭牆", 技術白皮書, 威播科技股份有限公司, 2002 08.