行政院國家科學委員會專題研究計畫　成果報告

以Linux為基礎的網路安全與頻寬管理閘道器之實作與研究

(II)

計畫類別：個別型計畫
計畫編號：NSC92-2213-E-009-125-
執行期間：92年08月01日至93年07月31日
執行單位：國立交通大學資訊科學學系(所)

計畫主持人：林盈達

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 93 年 12 月 8 日

一、計畫

計畫名稱	以 Linux 為基礎的網路安全與頻寬管理閘道器之實作與研究 (II)
計畫編號	NSC 92-2213-E-009-125-
主持人	林盈達 教授
執行機關	交通大學資訊科學系
執行期限	92/08/01 ~ 93/07/31

二、關鍵字

關鍵詞：病毒過濾，廣告過濾，網頁內容過濾，安全閘道器。
Keyword: anti-virus, anti-spam, content filter, and security gateway.

三、中英文摘要

針對日益嚴重的應用層內容安全問題，如病毒信廣告信，我們將Anti-Virus，Anti-Spam，Content Filter/Keyword三個內容過濾功能，與我們參加90學年度教育部通訊專題競賽獲獎之研究成果：7-in-1安全及QoS閘道器，整合而成支援內容安全之10-in-1網路閘道器，具管理集中，平台獨立等好處。此閘道器採用4-in-1 proxy架構，緊密結合四個套件的封包處理流程。套件包括(1)病毒過濾ClamAV, (2)廣告過濾SpamAssassian，(3) 網頁過濾DansGuardian，與(4)入侵偵測Snort。並去除與信件伺服器，網頁Cache套件Squid不必要的相依性。而Snort納入正常封包處理流程，除能偵測入侵，更能直接制止入侵(prevent intrusions)。Content Filter/Keyword方面，以N-gram演算法，找出東方語言關鍵字，提升東方語系網頁過濾準確度(69.6%->98.8%)。並提出Early Decision加速技術，提早阻止不適當瀏覽動作，縮短使用者等待時間為原來的1/4。

This work proposes a 10-in-1 content-aware security and QoS gateway for centralized management of content security problems, such as virus mail and spam. The gateway is derived from our previous achievement, 7-in-1 security and QoS gateway, which won the MOE project competition. Additionally, the 10-in-1 gateway owns three new content-aware functions, anti-virus, anti-spam, and content filter/keyword. In the gateway, the 4-in-1 proxy architecture unifies the packet processing flows of 4 proxies: (1) anti-virus *ClamAV*, (2) anti-spam *SpamAssiassian*, (3) content-filter *Dansguardian*, and (4) IDS *Snort*. The unified packet flow not only *doubles* the throughput but also enables the *Snort* to prevent, not just detect, intrusions. Also, the dependence of (1-3) on mail server and web cache *Squid* are removed. Besides the 4-in-1 proxy, the gateway increases the accuracy of filtering Eastern Web pages by selecting Eastern keywords with the N-gram algorithm. Finally, the novel *early decision* algorithm blocks illegal browses early and thus shortens the user latency by four times.

四、計畫緣由與目的

「網路安全」與「頻寬管理」已是企業上網的必備品。隨著Internet人口大增，網路用戶開始極度重視網路的安全性；也由於上網人數激增，對頻寬的需索無度造成某些重要的應用無法正常運作。為因應這些需求，許多方案已在市面上流通。 在91年的研究中，我們結合了Firewall, VPN, NAT, Bandwidth Management, IDS, URL Filter等七項實用的功能於一機，而完成7-in-1 security and QoS閘道器[1]。此閘道器以解決網路層所發生的安全及頻寬控制問題為主。上線使用與評比測試的結果證明了此閘道器的效率，穩固，及實用性。此項研究成果也在90學年度的教育部通訊專題競賽獲得優等獎，隨後技轉廠商，已產品化為D-Link DFL-1500及DFL-900, 於2003年度底開始銷售。

然而在使用此閘道器的過去這一年中，我們發現，網路層的安全問題雖已經解決，但卻因無法過濾病毒郵件，廣告信等發生於應用層的網路安全問題，而造成莫大的困擾。這類應用層的網路安全問題，其處理的對象直接以資料內容(content)為主，因此，又稱為內容安全(content security)問題。常見的解決策略，大都倚賴於直接在電腦或伺服器安裝相關軟體為主。

接下來我們針對病毒或廣告信件過濾，及不適當網頁過濾之目前常見解決方式及其缺點加以探討。

A. 郵件過濾

造成使用者困擾之電子郵件分為兩種：一為病毒信，一為廣告信。根據我們的觀察和測試，此類型解決方案，有三大問題。

1) 分散式阻隔，不易管理: 目前解決此問題之方法，大多倚賴在終端電腦安裝掃毒和過濾廣告信軟體。但此架構對企業而言，非常麻煩。往往並非每台個人電腦都能確實安裝相關掃毒軟體，及時更新病毒碼，最後依舊導致中毒。企業內部只要有一部電腦中毒，整個內部網路也隨之癱瘓。因此，我們建議應將此類工作集中至負責所有封包進出的閘道器來負責。管理人員僅需要設定閘道器的病毒碼或廣告特徵自動更新功能，便可達到同步保護整個企業內PC的目的。

2) 信件掃瞄效能: 根據我們的測試結果顯示，開放程式碼所提供的信件掃瞄套件(AMaVis [2] +ClamAV [3]+ SpamAssassian [4])，其效能並不良好，同時運作時，僅有2.85MB/sec的過濾速度。由於目前所流行的病毒，其發作時，常以狂送病毒信來達到散播目的。若掃瞄病毒的速度不夠快，太量病毒信將累積在伺服器上，而導致上傳或下載信件嚴重延遲。顯然，效能問題是不可忽視的。因此本研究藉由底下兩個方式，來找出瓶頸點。一是剖析系統運作狀況，以釐清掃描信件處理步驟。另一是輔以白箱測試，量測各處理步驟所佔用時間比例。結果發現，由於原本這些開放程式碼套件，在運作時皆須互相搭配，才能達到功能，例如信件過濾套件AMaVis需搭配mail server。因此一封信件便必須在很多個行程中交互傳遞或是待掃瞄內容得在Kernel/User Space間交換，而影響效率。因此我們才會提出一個統一的proxy

架構應用於內容安全閘道器上，以單一化AV, AS, CF, IDS四個功能的封包處理流程，減少IPC的資料量，以增加處理速度。

3) 千變萬化的廣告信: 廣告信內容過於多樣性，若欲藉由電腦判定是否為廣告信，有相當的困難。部分商業軟體，甚至僅以一黑名單比對信件來源，確定是否為廣告信，可想而知，此方法之漏擋情況必然非常嚴重。開放程式碼所提供的Anti-Spam套件(AMaVis+ SpamAssassian)，採取特徵值計分累積超過某一門檻值的方式來判定廣告信，雖有初步效果，但根據我們一系列的測試及果顯示，仍有誤擋電子報,或是30%漏擋的問題。此種累積計分有兩大問題。

特徵值分數給定：各特徵值與分數，其實無絕對的關係。舉例來說，原本該套件判斷廣告信的一個特徵便是非西方語系信件。當信件內容為非西方語系時，給予很高的計分，認定為廣告信。但顯然，這樣的配分，並不適用於東方語系的國家。

門檻值設定：當一信件計分累積到門檻值時，該套件便認定為廣告信，但問題是，該門檻值之設定，並無一定準則。當此門檻值設定過高，會發生漏擋。反之當門檻值過低，則造成誤擋。這個問題，在IV.A節有進一步的實驗數據來呈現。

B. 網頁過濾

1) 以URL list阻擋的問題(速度與漏擋): 對父母來說，通常擔心孩童瀏覽色情或暴力內容的網頁。而對企業雇主來說，則不希望員工瀏覽股票或其他影響工作效率的網頁。因此，一部提供網頁內容過濾功能的閘道器，是迫切需要的產品。目前市面上，常見的內容過濾方式，大都僅以比對網址，來達到過濾不恰當的網頁的目的。此「網址比對過濾」有兩個主要的問題。

龐大的Blocking URL List：在Internet上，色情或其他不適當網頁的數量，是非常驚人的。若僅藉由網址比對來達到過濾目的，那儲存一個龐大的URL清單是無可避免的。此外龐大的資料庫，已經嚴重影響比對的速度，根據IV.A節實驗數據顯示，網址過濾速度(350KB/s)，甚至於比掃瞄整個回傳網頁的過濾速度(440KB/s)還要慢。

未即時更新而導致漏擋：由於不適當網頁數量，往往增加迅速。而要發現不適當的網頁，又必須依靠人工的方式檢舉判定。因此若不能及時發現該網頁，並加以更新資料庫，便會造成該網頁的漏擋。

2) 以網頁文字阻擋的問題： 速度與語言:針對以URL比對問題，目前開放程式碼套件DansGuardian [5]，進一步提供，以關鍵字掃瞄傳回網頁的方式，過濾不適當的瀏覽動作。此方法實際檢查內容，因此較能準確阻隔。此外由於儲存關鍵字所需之空間，相較於Blocking URL List，不僅數量較少，且增長速度也很緩慢，因此是一個比較好的方法。但此方法仍然存在兩個問題：

(i) *未能精準過濾中文網頁:*由於原有套件為英語系國家開發，以處理拼音語系的關鍵字為主。拼音語系單字本身便已具有詞的特性和意義。若比較東方語系來說，所謂的詞往往必須由一個以上的字來組成，這時候在比對的演算法上是需要若干修正。此外就關鍵字資料庫來說，原有之資料庫，必然無法分辨中文內容的網頁。因此重新建立中文關鍵字資料庫是有其必要的。在本研究中，我們應用N-gram演算法[7]，在一百個色情網站的網頁中，分別以遞增長度的詞，來統計其出現頻率，最後找出頻率高於門檻值的不定長度關鍵詞。

(ii) *網頁內容傳遞緩慢，增加使用者等待時間:*當以Request所要求擷取的網址作為比對目標，可在使用者發出Request後很短的時間內，便予以回應。但當以Response內容為過濾目標時，Response內容龐大，又得從遠端的主機傳送回來，整體接收時間很久，增加使用者瀏覽一個新網頁或被告知不可瀏覽回應的時間。因此針對這問題，我們提出Early Decision技術，針對傳送中的網頁，其最可能歸屬類的得分明顯高過其他類別時，便可逕行決定。若確定不允擷取，則及早回應使用者，並同時停止後續內容傳送，節省網路資源。

五. 研究方法

甲.主要貢獻

本研究利用開放式軟體作為系統各功能開發的基礎，有效節省成本及縮短開發時間。但正如IV.A節中黑白箱測試結果顯示，此類軟體具有效能不彰的問題。我們創新的重點在於根據測試結果，挖掘效能瓶頸，研發新技術，以改善效率問題。因此貢獻可分為自行研發新技術，和系統整合應用兩方面來介紹。

A.自行研發技術方面：

1) 新 4-in-1 Proxy 架構 (正申請專利): 透過白箱測試的結果顯示，多套 proxy 功能之開放程式碼套件,如 AMaVIS+ Clam-AV+ AS 或 DansGuardian,的架構效率不彰，是因為這些功能的運作需要配合其他套件。如 AMaVIS 需搭配 Mail Server, DansGuardian 需搭配提供 web cache 的 squid 套件。由於 mail server 及 squid 並不是一個安全閘道器所需要的套件，這導致信件內容很無意義的在數個程式行程中傳遞(IPC 問題)以及 Kernel/User Space 間轉換，這傳遞存在很大的延遲(詳細原因分析請見 3.A 節)。這個研究的貢獻，除以實際測試結果找出此些瓶頸點外，我們進一步提出 4-in-1 Proxy 架構，單一化 ClamAV, SpamAssassian, DansGuardian, Snort [6]四個過濾套

件的封包處理流程。去除上述相依性，使資料交換皆在單一行程中完成，此外，避免原本負責 IDS 功能的 Snort 自行複製封包，二次合併所造成的系統資源浪費。當然也由於此單一化的過程，Snort 被納入標準封包處理流程中，不僅可以偵測入侵攻擊，更可以達到阻擋的效果(Intrusion Prevention)。現在將此架構申請台灣與美國的專利中。

2) 東方語系關鍵字支援: 在原有開放程式碼 DansGuardian 套件中，雖已經能以片語方式比對網頁內容，但所要比對的片語，皆為西方語系。本研究，蒐集將近百餘個色情網站的網頁，應用 N-gram 演算法，逐一統計長度漸增的片語。按照該片語出現的頻率，來選出適當的關鍵字。相較於以人工直覺的方式來選出片語，更能挑選出具代表性的詞類。舉例來說，人工直覺可能會認為美女，美腿等字眼容易出現在相關情色網站。但結果根據我們 N-Gram 的統計，可能常出現的字眼是「未滿十八歲」，這樣完全跟色情無關的字。在增加了中文關鍵字後，根據實驗結果顯示，其中文色情網頁的阻擋率，已然從 69.6%增加到 97.2%。

3) Early Decision 技術(已申請專利): 以關鍵字搜尋回應網頁內容的方式，對系統資源來說，雖已經比大量的網址比對來的有效率和精準。但單就使用者感受到的反應時間來說，反而因為要等內容傳回後才能決定，而來的比較久。本研究提出 Early Decision 的技術，針對傳送中的網頁，當其最可能歸屬類別的得分明顯高過其他類別時，便及早決策。如此對於允許通過之網頁，使用者可以先行瀏覽已經傳到的內容。而對於不予允許的網頁，使用者可及早知道，此外閘道器亦可以停止後續內容的傳送，避免網路資源的浪費。根據實驗結果顯示，Early Decision 技術提升處理效能達 4 倍，縮短延遲時間為 1/3。此方法已申請台灣與美國的專利。

B.系統整合應用方面

在系統整合方面，我們延續參加 90 學年度教育部通訊競賽獲獎作品：7-in-1 Security and QoS 閘道器的成果，新增加三項內容過濾功能 Anti-Virus, Anti-Spam, Content Filter /Keyword， 而成為能處理網路層和應用層安全問題之 10-in-1 安全及 QoS 閘道器。

就各單獨功能來說，在 Anti-Virus 部分，統一於閘道口阻隔一切的病毒信攻擊，除避免 Mail Server 癱瘓外，並避免某些懶惰的使用者，因未及時更新病毒碼，中毒並癱瘓公司內部網路。Anti-Spam 方面，不論對員工或任何一般使用者，都是目前急需的功能，已不需贅述。而 Content Filter 更是目前各企業雇主及家庭中的父母的需求，以往產品雖有此功能，但純粹只靠網址比對，若資料庫未能即時加入該比對網址，往往不能有效過濾。此外龐大的阻擋網址列表，根據我們 IV.A 節的實驗結果顯示，嚴重導致比對速率下降。而在我們 10-in-1 的閘道器上，整合的乃是支援 Keyword 比對能力的過濾器，此外因為我們以 N-gram 演算法擷取東方語系關鍵字，對於比對中文網頁的準確率(98.8%)，更是一般目前不論產品或軟體，所不能比擬。

就整合後成本價值來看，對企業來說，僅需要購買單一台安全閘道器，即可快速避免及解決從網路層到應用層所有攻擊及需求，不僅十分經濟，尚且十分容易安裝和設定。完全不需要考慮到多台網路設備間，設定相互矛盾時，導致網路不通的問題。這一切，都已經在 10-in-1 閘道器整合之初，已經解決。

雖然本研究已減低了 system ovehead，但在 signature matching 方面，仍有改善的空間，因此未來一個主要的改良方向便是設計 signature matching 演算法及硬體加速。此外我們也計畫將 application QoS 的相關功能加入本整合閘道器中。

乙. 設計原理分析

A. Proxy 架構分析

1) 原有系統架構: 如圖0所示，在舊架構當中，除了Snort是用BPF將封包從kernel複製上來之外每個application都是聽特定的port，因此這樣一來，當有traffic需要被兩個application檢查的時候，封包就會被重複的處理。另外從traffic flow來看，因為這些application除了IDS之外都需要與其他的application搭配使用，其中DansGuardian需要與Squid，而AMaVis則需要與Mail Server和SpamAssassian模組以及ClamAV，所以會另外產生多次的user/kernel interaction與inter-process communication。這些application會重複的處理封包以及產生多次的user/kernel interaction和IPC，對於系統是一個很大的overhead，所以在效能上無法有很好的表現。

2) 新4-in-1 Proxy架構: 圖0是我們所提出將原本4個獨立的Proxy模組(Clam AV, SpamAssassian, DansGuardian, 及Snort)緊密整合以後的架構，這樣的新架構，有3個好處。

(i) *加值功能嵌入至proxy:* 去除DansGuardian與安全閘道器中不需要的Squid套件之相依性，而改以與我們自行撰寫之web proxy相接，達到系統瘦身的效果。

(ii) *以shared lib偵測並阻擋攻擊*: snort套件從原本一個單獨的process而簡化為一個shared lib，供web及mail處理過程呼叫使用。如此原先在舊架構中，封包需要被複製一份來自行重組和檢查的過程，即可被省略，避免同一封包在系統中重複處理。更重要的是，這樣的調整，使得snort可以提供Intrusion Prevention的功能。

(iii) *減少IPC:* 收到的信件，將直接傳給AMaVis處理，不需轉手於Mail Server。去除掉從mail server將信件透過IPC導入AMaVis所需要的時間，而能增快效能。根據IV.A節實驗結果分析，IPC時間約佔整體信件處理時間將近47%。因此降低IPC的交換量，改進效果是可期的。

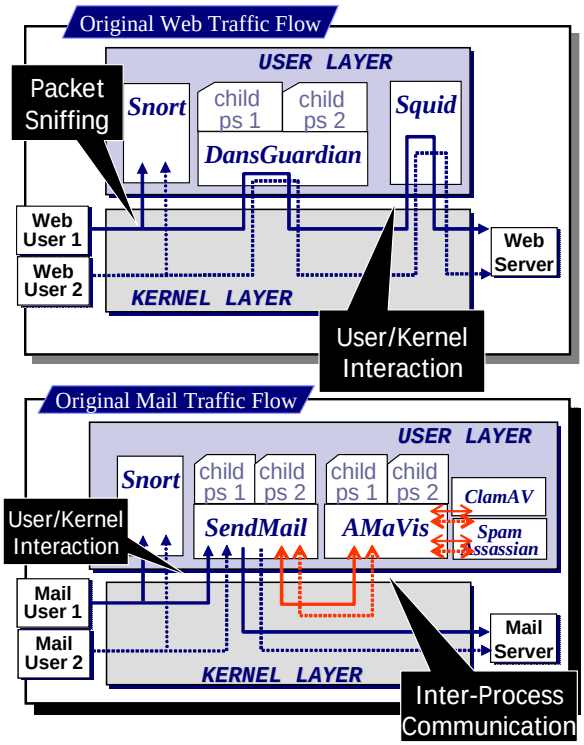

圖 0:網頁與信件流過閘道器之原先內部流程(改進前)

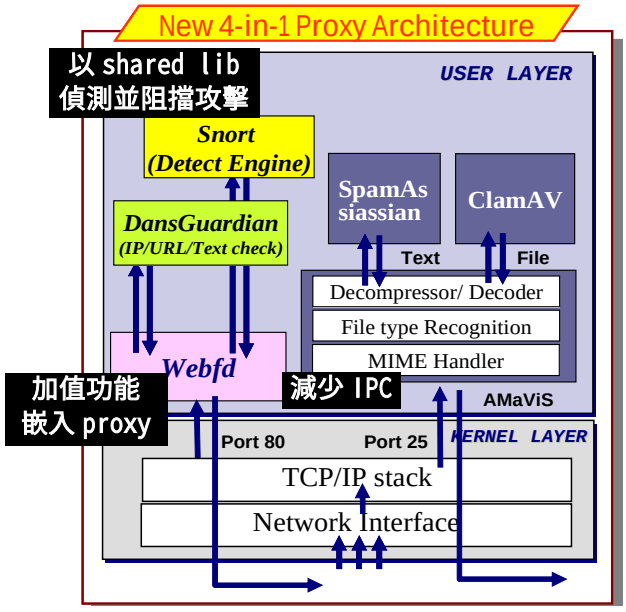

圖 0: 單一 packet flow 的 4-in-1 Proxy 架構

B. 東方語言過濾

為了能讓內容過濾器能夠處理中文的網頁內容，我們改進步驟如圖 0 所示。Stage 1 是收集夠多同一類型網頁內容的樣本（如色情類、股票類等)。State 2 則包含使用 N-gram 的演算法來統計文件內的關鍵字，N-gram 的演算法主要的方式是使用 2 個字、3 個字.....直到 N 個字的統計，假設有一關鍵字如"大盤成交量"，用 2-gram 處理時會得到"大盤"、"盤成"、"成交"、"交量"等四個關鍵字，再用 3-gram 處理之後得到"大盤成"、"盤成交"、"成交量"等三個關鍵字，以此類推，到 5-gram 時，便會得到"大盤成交量"這個關鍵字。

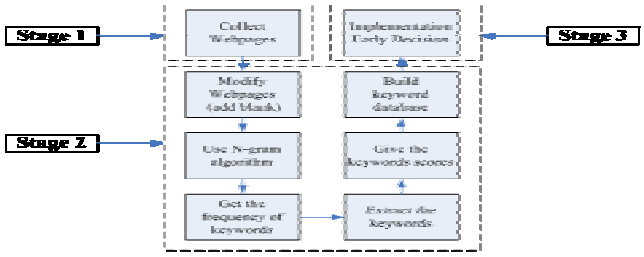

圖 0: 東方語系關鍵字擷取步驟

應用 N-gram 時會遇到以下三個問題：(1) N 要取多少，(2) 關鍵字的擷取及分數，(3) stop words（extremely common words)。第一個問題的 N 值，可在訓練階段觀察關鍵字出現的頻率而得知，例如當出現頻率已低於 5 則停止，而此時所取到的關鍵字長度即為我們所要的 N 值。至於第二個問題，關鍵字的擷取及配分，可根據三項特點來決定： frequency, breadth, length。由於是使用取出的關鍵字再來判別網頁是屬於那一類，所以此關鍵字出現的頻率是第一要素，如果出現的頻率太低，則沒有分類的價值。其次是根據廣度來決定是否要選取，因為如果在同一類型的網頁中，某關鍵字只出現在少數的網頁中，則代表廣度不夠，最後則是針對長度來探討，因為較長的關鍵字較能決定其所屬的類別，如"交通大"這個關鍵字，與"交通大學"這個關鍵字相較起來，後者可得知是學校，而前者較不能決定，因此，較長的關鍵字我們給予較高的分數，在我們的實驗中，我們是將關鍵字出現的頻率當成其分數的依據或是相等。最後一步驟就是將這些得到的關鍵字加入內容過濾器裡的關鍵字比對資料庫中並給予關鍵字分數，關鍵字的分數是利用 N-gram 處理時得到的出現頻率來決定。

至於第三個問題 stop words，亦可以利用另一組正常的網頁進行以上步驟的方式產生出來的關鍵字，如果兩類都出現此關鍵字則刪除這個關鍵字，因為 stop words 必然出現在所有類別的網頁，所以使用此方式便可刪除一定數量的 stop words，減少人工篩選時的負擔，經過前面的刪除之後，最後再使用人工篩選的方式確認取出的關鍵字是否有義意。

C. Early Decision技術

如圖 0 所示，此技術分為 Early blocking 和 Early bypassing 兩部份，由於傳統內容過濾器在處理網頁內容時，是累計網頁內容裡出現的關鍵字的分數，整

份文件統計完之後，再決定是否該阻擋此份文件（不讓用戶端瀏覽）或讓其通過（讓用戶端瀏覽）。由於是處理完整份文件，所以增加用戶端等待時間。而我們所提出的 Early Decision 技術則不需要完全看完整份文件，因此可加快處理網頁的速度，也可讓用戶端不需浪費許多時間在網頁文件的等待。作法是採用門檻值的方式，在 Early blocking 方面，在統計網頁分數時，會找最高分數的兩類來計算其分數比例，如果最高的值與次高的值差距比率達到阻擋門檻值的話，便可以直接將此網頁阻擋，不需要再處理完整份文件；而在 Early bypassing 方面，作法上大致與 Early blocking 的作法相同，也是統計各類別關鍵字的差距比率，取最高與次高的分數來計算，如果兩者的關鍵字間距大於間距門檻值，且兩者分數小於阻擋門檻值的話，便可以讓此網頁通過，舉例來說，如果此網頁屬於股票類，在統計分數時，股票類關鍵字的分數必定高於其它類關鍵字的分數，當差距到一個門檻值時，便可以猜測此份文件應該屬於那一類的網頁內容，如果差距不明顯及關鍵字的間距大於間距門檻值時，便讓它通過。

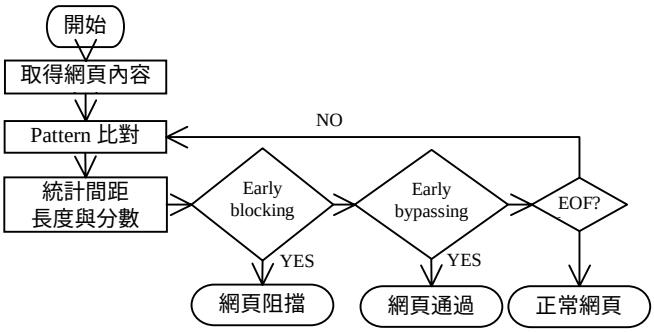

圖 0: Early decision 功能之執行流程圖

D. 新10-in-1 閘道器之操作介面

有鑑於原本開放程式碼套件，並未提供友善的操作介面，在我們整合改進這些套件之餘，也設計了一個透過WEB可以進行相關網路設定的介面。圖1是此閘道器登入後的畫面，從左邊的選單可以看出，此閘道器除基本閘道器功能外，所提供的進階功能，如Web Filter, Mail Filter等。圖2則是Mail Filter的操作介面而圖3則是Web Filter的操作介面。


圖1: 支援內容安全之10-in-1網路閘道器歡迎畫面

圖2: 10-in-1 網路閘道器之AS畫面

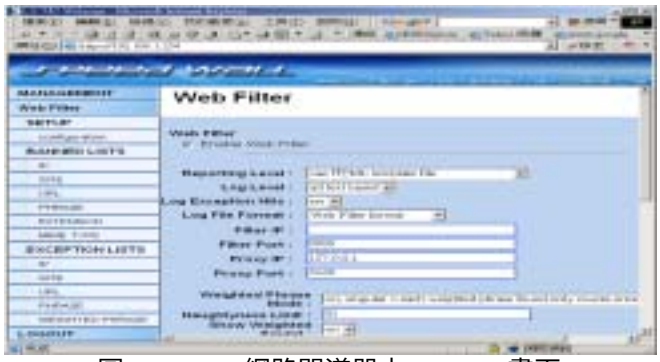

圖3: 10-in-1 網路閘道器之web filter畫面

六、結果與討論

A. 郵件掃瞄

1)效能 （黑箱測試）

在這項測試中，我們分別使用相同以及不同的郵件做測試。圖 3 顯示，在 Proxy 的模式當中，mail server 只幫我們作送信的工作，throughput 可以到 25 Mbps，可是一旦開啟 AMaVis 之後，throughput 只剩下 4.4 Mbps，而開啟 Anti-Virus 以及 Anti-Spam 之後，throughput 甚至掉到 2.85 Mbps。而送相同以及不同的郵件會造成差異，是因為當 AMaVis 在做掃毒的動作的時候，如果發現這封郵件是掃過的郵件的話，那就不再重複掃毒。

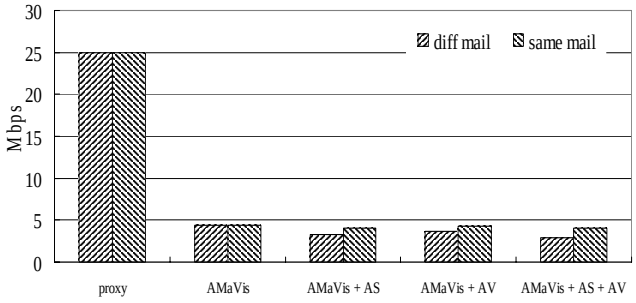

圖3: AMaVis,AS,AV套件功能開啟之效能影響結果圖

2) SPAM過濾準度 （黑箱測試）

(i) 誤擋測試: 圖3為信件的誤擋測試，測試抽樣採用100封使用者訂閱的電子報，若門檻設定為5分，可以看出誤擋的情形很嚴重，顯然由於電子報與廣告信的特徵相近，而導致目前方式無法正確辨認。圖3之測試抽樣改採100封朋友彼此傳閱的信件，若門

檻設定為5分，可以看出有一封信件超過門檻值，針對此信深入追蹤，發現其內容只有一張圖，與廣告信特徵類似，所以被阻擋，因此本誤擋尚屬合理。

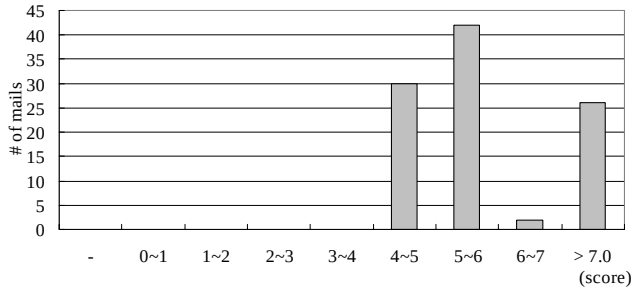

圖 3: 100 封電子報之計分結果

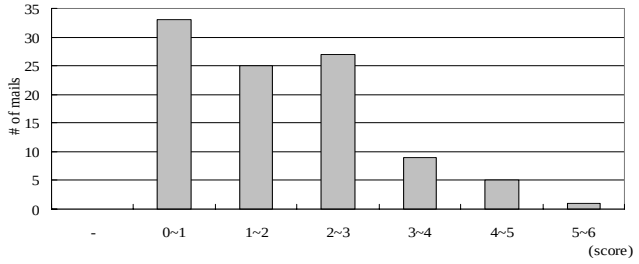

圖 3: 100 封朋友間轉寄信件之計分結果

(ii) 漏擋測試 (黑箱測試): 信件的漏擋測試抽樣採用1000封垃圾信，圖3為測試後之結果。由於AMaVis內訂有假設廣告信小於64K的設定，因此結果以兩條線分別呈現此設定是否開啟的差異。當default開啟時，可以看出有30%的廣告信(方塊線中N/A的點)沒有分數，而造成有44%的信件小於五分的門檻值，而導致廣告信的漏擋。既使將此限制去掉時，則漏擋率(<5分)仍有30%。

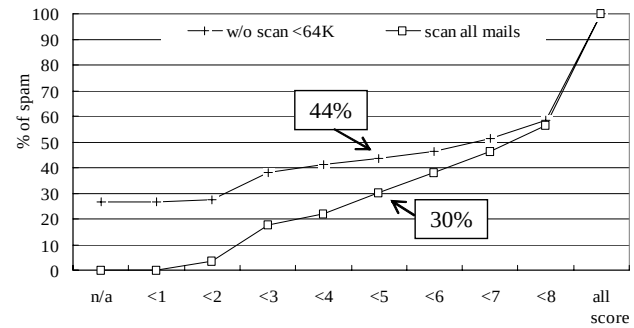

圖 3: 1000 封廣告信的漏擋測試結果

3).瓶頸點確定 （白箱測試）

在這項測試中我們持續輪流以十封不同夾帶大檔案的信件讓 AMaVis 來掃描，測試各個部份分別要花多少時間。選擇用十封不同信件輪流寄送，是為了排除 AMaVis 中 Cache 功能影響真實測試結果。圖 3 顯示，一封信真正被掃瞄的時間僅有 AMaVis 的 491.4ms 加上 AVScan 的 2116.6ms，僅佔全部時間 6164.6ms 的 42%。而其他花在與 mail server 行程間的資料交換 (enqueue, smtp receive, smtp forward) 占 2900/6164=47%。此顯示除一味的改進掃瞄演算法之外，系統本身效能的改進(如 IPC 間的效能)，也是十分重要的。

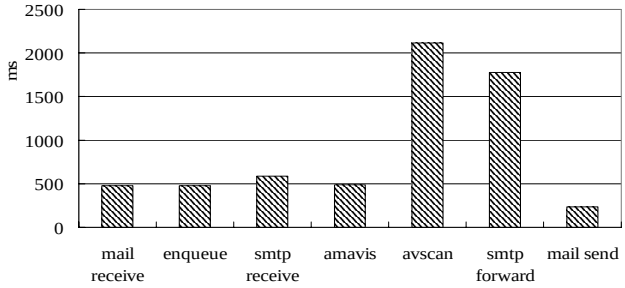

圖3:AMaVis 內部處理信件各流程之執行時間

B. 網頁內容過濾

1) 網頁內容過濾效能 (黑箱測試): 圖3及圖3的數據是模擬10~80個用戶端至網頁伺服器上取得5KB大小的網頁，其中None是指完全不開啟任何檢查項目，URL是指只開啟URL資料庫的檢查，URL Keyword是開啟檢查有關URL Keyword的檢查，Content是開啟網頁內容文字的檢查，ALL是指以上三項檢查全部打開。圖3顯示，既使僅啟動DG但不開啟其任何功能，其處理速度便只有550KBps (4.4Mbps)。另外，比較URL與Content的線可以得知，龐大的URL名單比對速率(350KBps)是比content比對速率（440KBps）還糟糕的，顯然以關鍵字過濾content，當URL名單日益龐大時，是相對來的有效率的。而圖3則顯示每秒DG所能處理的使用者需求數，大約從所有功能皆不開啟的105 requests到所有功能皆使用的60 requests。

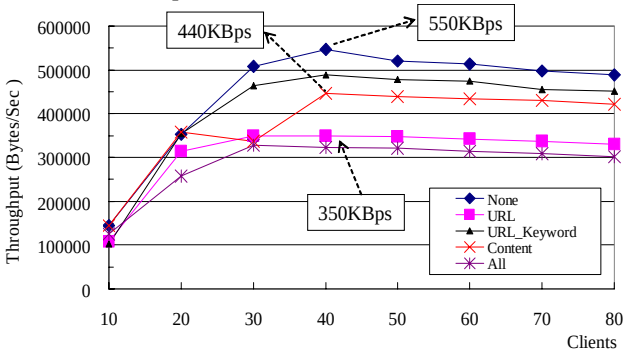

圖3: DansGuardian三種功能開啟與否影響網頁過濾效能差異

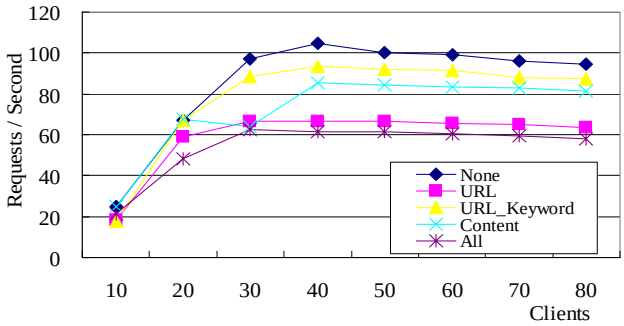

圖3: DansGuardian三種功能開啟與否影響處理需求速度差異

2) 瓶頸點確定(白箱測試): 測試結果顯示，在Server端回應給Client所花費的時間中，Content (網頁內容過濾) 比對時，幾乎佔了所有的時間(99.72%)，由此可知內容過濾器的主要瓶頸還是在處理Content的部份，必須將整個網頁內容掃過一遍，然後以累積分數來分類，看是否要阻擋。在確定瓶頸點在這邊後，我們因此而提出Early Decision的概念，解決這個問題，以期能縮短這部分的處理時間。

3) Early Decision改進效能(黑白箱測試): 在改進效能的測試部份，我們使用Web Bench 5.0，以7台Pentium III 1GHZ的電腦當作Clients，每台模擬10個Clients，而Web Server則是用Pentium 4 1.5GHZ，在其上放置了40KB的網頁，每個Client分別經由content filter對Web Server發出Request。從圖3及圖3黑箱測試結果顯示，具有Early Decision與沒有Early Decision的效能與Request的數量改進將近3倍。

而內部測試方面，latency 是最為重要的量測重點，因為過長的 latency 將會使 Content filter 的實用性大大降低。測試的方式是在程式中加入一些記錄時間的程式碼，經過連續處理 100 個網頁資料之後，分別記錄有 Early Decision 與沒有 Early Decision 兩種處理所花費的平均時間，但由於 Early Decision 的兩部份（Early Blocking 與 Early Bypassing）在性質上不一樣，所以不能用相同的網頁進行測試，因此在 Early Blocking 方面，我們是使用實際的色情網頁，網頁大小分別為 1KB, 6KB, 18KB, 29KB，在 Early Bypassing 方面，則是使用正常該通過的網頁，分別為 Google（4.12KB)、NCTU (20.1KB) 及 PCHOME (35.6KB)。從圖 3 可發現，有 Early Blocking 的 Latency 比沒有 Early Blocking 快將近 4 倍。而從圖 3 亦可發現，有 Early Bypassing 與沒有 Early Bypassing 在 Latency 上差距則不一定，主要是因為掃描到文章的每個階段（掃描百分比）是否有低於該階段分數的門檻值，如果低於門檻值的話，便可以儘早判斷此網頁是否屬於該阻擋的網頁。

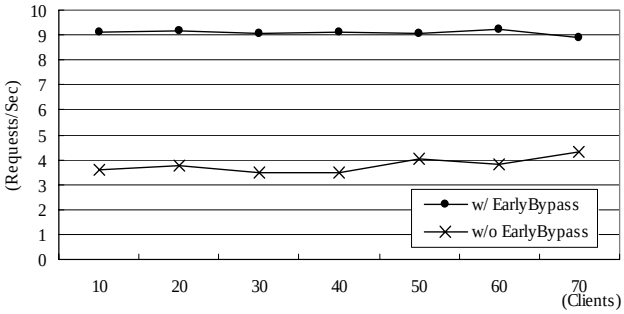

圖 3: Early Decision 使用前後之 Request 處理數量比較

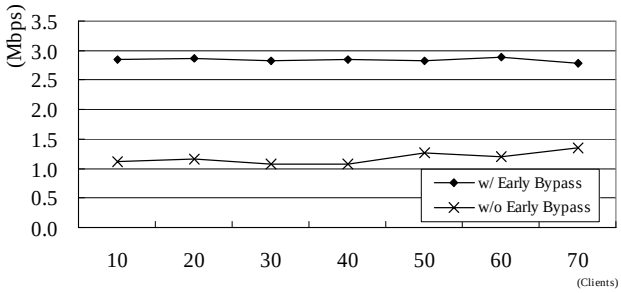

圖 3: Early Decision 使用前後之處理效能速度差異

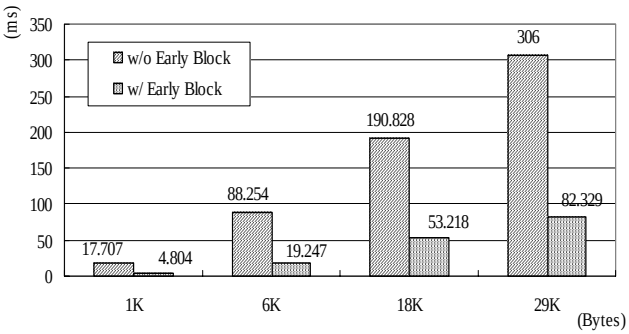

圖 3: Early Blocking 使用前後不同網頁大小之使用者延遲比較

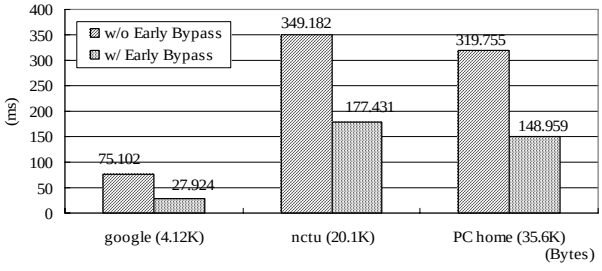

圖 3: Early Bypass 使用前後不同網頁大小之使用者延遲比較

4) 準度改善 (黑箱測試)

表1蒐集國內250個色情網站與國外250個色情網站進行的準確度測試，分別開啟URL、URL Keyword、Content Keyword等進行測試，其中Content Keyword部份是加入中文關鍵字之前與之後的數據。光看國外網站阻隔率，可以發現，URL+Content Keyword過濾了大部分的不適當的國外網站(98%)。另外從中文網站過濾結果顯示，在增加了中文關鍵字

後，光是Content Keyword的阻擋效能，便從69.6提升到97.2。而三樣功能全部使用的結果，阻擋率則可提升到98.8%。

七、結論

在這研究中，我們將三個實用的功能, Anti-Virus (AV), Anti-Spam (AS), 及 Content Filter (CF)/Keyword，整合進我們91年研究成果7-in-1安全及QoS閘道器中，而成為能同時處理網路層和應用層安全問題的10-in-1 內容安全閘道器。藉由對單純安裝完套件(架構未改進前)的閘道器進行黑箱測試(External Benchmark)顯示，開放程式碼套件如AMaVis的郵件過濾的速度僅有2.85Mbps而，Dansguardian的網頁過濾速度也僅有4.177Mbps，遠低於原本網路層安全檢查速度，一般可接近線速(wire speed)。白箱測試結果分析得知，此兩個套件在原本的設計，分別需搭配對安全閘道器來說，多餘的Mail Server及squid(提供Web cache功能)來運作，而導致一封信件或一個瀏覽動作，需在多個行程中無意義的傳遞處理，嚴重影響效能。此外Kernel/User Space間的資料傳遞瓶頸也是不可忽略的。針對這些問題，我們提出了4-in-1 proxy架構，單一化AV, AS, CF, IDS四種過濾工作的封包流程，去除AMaVis與mail server的相依性，DansGuardian與squid的相依性，Snort複製封包及二次合併的系統資源浪費。此單一化，也使Snort除提供intrusion detection更能直接阻擋攻擊，而達到intrusion prevention的功能。

另外實驗顯示DansGuardian中content keyword的阻擋功能，對於東方語系網頁阻擋率僅有69.6%左右。因此我們以N-gram演算法，從一百個色情網站的網頁中，統計出不定長度的關鍵字，加入關鍵字資料庫中，明顯將阻擋率提升到97.2%。此舉除了改進CF的阻擋率外，更驗證以N-gram找尋關鍵字的能力。另外由於以content為過濾對象，在原本的DansGuardian中需判讀全部的網頁才能決定是否過濾，嚴重延遲使用者等待時間。我們提出Early Decision的方法，對於傳送中的網頁，如果其最可能歸屬類的分數已經明顯高於其他歸屬類型，便可逕行判定該網頁屬於該最高歸屬類型之網頁，而不必要判讀完全部的內容。實驗結果顯示，Early Decision可提高處理效能將近3倍，使用者等待時間縮短為1/4。

八、計畫自評

在計畫成果自評部份，請就研究內容與原計畫相符程度、達成預期目標情況、研究成果的學術或應用價值、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。

若有與執行本計畫相關的著作、專利、技術報告、或學生畢業論文等，請在參考文獻內註明之，俾可供進一步查考。亦可將相關內容當作本報告附件，繳送國科會結案。

九、參考文獻

[1] Ying-Dar Lin, Huan-Yun Wei, Shao-Tang Yu, “Building an Integrated Security Gateway: Mechanisms, Performance Evaluation, Implementation, and Research Issues,” IEEE Communication Surveys and Tutorials, Vol.4, No.1, third quarter, 2002.

[2] AMaVis, http://www.amavis.org/

[3] ClamAV, http://www.clamav.net/

[4] SpamAssassian, http://news.spamassassin.org/

[5] DansGuardian, http://dansguardian.org/

[6] Snort, http://www.snort.org/

[7] Fuchun Peng, Dale Schuurmans, “Combining Naïve Bayes and n-Gram Language Models for Text Classification,” The 25th European Conference on Information Retrieval Research (ECIR), Dec. 2003.

表1: 增加中文關鍵字前後過濾國內外網站之準確度測試結果比較

URL	URL Keyword	Content Keyword	Domestic (250)		Overseas (250)	
			Blocked pages	Blocked ratio	Blocked pages	Blocked ratio
◎			159	63.60%	241	96.40%
	◎		4	1.60%	41	16.40%
		◎	174→243	69.6%→97.2%	226	90.40%
◎	◎		159	63.60%	241	96.40%
◎		◎	218→247	87.2%→98.8%	245	98.00%
	◎	◎	175→243	70.0%→97.2%	227	90.80%
◎	◎	◎	218→247	87.2%→98.8%	245	98.00%