# 網路封包監聽技術應用於網站使用者行為分析

蘇建郡*　林明達**　施志鴻**
南台科技大學計網中心 *　南台科技大學資訊管理所 **
ccsu@mail.stut.edu.tw *, m9190210@email4.stut.edu.tw **,
m9090104@email3.stut.edu.tw **

## 摘要

近年來網際網路迅速發展，網路帶給人們無遠弗屆的感覺，網路的使用率也跟隨著寬頻的推動漸漸增長，所以藉由網路封包監聽技術來分析網路流量也越來越被重視，因此藉由對於網際網路運作原理與通訊協定的認知，我們利用此技術來分析網路最常用的 Web 功能，分析此封包使用的通訊協定（HTTP - HyperText Transfer Protocol），紀錄使用者透過瀏覽器來瀏覽 Web Server 的相關網站的資訊，再加以匯整統計，得知此群組網站的熱門網頁的排行板，最常下載檔案排行等相關資訊，更能真實重現訪客的瀏覽過程，並將所得資料直接存於資料庫中取代 log 方式，以方便管理者能更快速的查詢。

**關鍵詞：**網路封包監聽、通訊協定、HTTP

## Abstract

The internet has been developed widely using in recent years, and it changes the world a lot. The using-rate of the internet keep rising while the broadband network keep growing. To analyze the traffic flow by network monitoring technology is more important than ever. Through the principle of the internet and the communication protocol, we analyze the HTTP of this packet and recording the information of web sites that user browsed by browser with these data we know the information of top web sites and top download files further more we understand the user's path on the internet. We kept those information in database for quicker search than log data for managers.

**Keywords：**network traffic monitoring, communication protocol, HTTP.

## 前言

網際網路的進步，驅使愈來愈多使用者與企業透過網路來進行許多行為，對於離島地區，不管是要推廣觀光活動或者進行商業交易，網路更能發揮其特性，因此如何透過網路技術應用在生活上變成很重要的一環，近年來電子商務與網路行銷持續被推廣，B2B、B2C、C2C 的詞彙不斷出現，強調透過網際網路的好處，但是在這前提下，必需以技術當做基礎，才能更進一步的運用，以網站來講，了解使用者瀏覽自己的網站行為，一直是管理者所興趣的，目前大多數的做法是透過計數器的方式，來代表此網站的人氣指數，但是這個值並不是絕對的，且所能代表的資訊是相當少的，因此我們針對這點，利用我們對與網際網路運動原理的了解，結合網站運作所使用的通訊協定，設計了一套針對 HTTP 封包分析的系統。

此系統不但能針對拜訪者的來源 IP 做記錄，更能針對拜訪者瀏覽過哪些網頁、下載過哪些檔案、拜訪時間等等資訊做記錄，以供管理者做即時查詢或者事後分析，另外此系統並不是只針對一個網站做監控行為，只要所架設的網站符合我們所說明的環境，均能做記錄、分析。

## 系統環境

本系統主要以 Open Source 的 Linux Red Hat 7.3 做為開發的平台，運用封包監聽的技術將取得的資訊存入資料庫，來取代 log 的方式，並分析封

包所包含的資訊，運用此方式好處在於無需更改原電子商務所架設的系統與網頁，只需放置此系統於網站對外出口主機即可，如下圖（圖 1 系統環境 方法一），或者透過 Router 將網站所接收到的封包 mirror 至分析系統主機（圖 1 系統環境 方法二）即可，完全不影響原有網路效能。

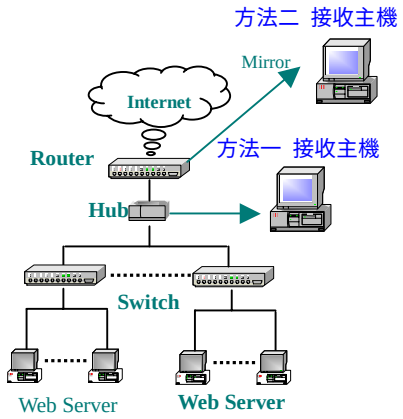



圖 1 系統環境

## 知識原理

封包監聽系統主要以 Linux 為開發平台，以 Linux C 為開發程式語言，利用執行緒多工原理及共享記憶體的方式，並設定網路卡為接收全部封包模式(Promiscuous)[7]，來做封包的接收及處理，擷取屬於 IPV4 的封包。

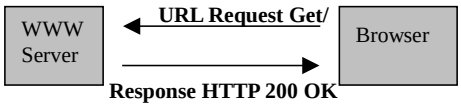
封包監聽系統最主要針對 IP 層(Layer3)及傳輸層(Layer4)、應用層(Layer7)來做封包資料的分析，藉由 IP 層、傳輸層最主要得知使用者從何處來、下載量多大、是否對網站主機造成過大負荷，在應用層方面，最主要得知使用者抓取哪些檔案，瀏覽過哪些網頁、存取過那些圖片檔等相關資訊，藉由此方面資料來分析，詳細說明如下。

**IP 層(Layer 3)：**取得使用者來源 IP、目的 IP，可得知使用者來自何處，由封包欄位 total length 累計得知下載檔案大小資訊，如果針對來源 IP 為此電子商務網站的 IP，而目的 IP 為本公司以外的 IP，表示是使用者連到此網站的記錄，藉此可統計使用者對此網站的網路頻寬使用量為何。

**傳輸層(Layer 4)：**取得封包的 Port Number，得知使用者進入電子商務網站所下載的檔案到底是透過 FTP 或 WWW 下載等相關資訊，電子商務網站可藉此數據來調整內部檔案存放下載方式，如欲得知此網站 FTP 流量多大，可由 Port 為 20、21，要得知 WWW 則為 80，如欲得知透過 Email 流量，內送郵件 POP3 可由 110 Port 得知，外寄郵件 SMTP 可由 25 Port 得知。

**應用層(Layer 7)：**主要以 HTTP （HyperText Transfer Protocol）[2]為主，HTTP 運作方式(如圖 2)，Client 端瀏覽器會利用 HTTP 至網站伺服器中請求資源，此為 Request 封包(封包內容如圖 3)，而在 Request 封包皆有一個 URL 的獨特識別字，如 http://www.mis.e du.tw/Top_Files/sethome.GIF，擷取出封包內容查看其封包格式，顯示 Get /Top_Files/sethom e.GIF，所以只要判斷封包的檔頭資訊是 Get，取出其 URL (如上例只要取出 /Top_Files/setho me.GIF)，利用此特性就可知道訪客來此網站曾下載過什麼檔案，瀏覽過什麼網頁（如 Get /index.htm）等多種相關資訊，甚至可知是否有使用者透過 Unicode 安全漏洞[1]，利用 URL 方式執行 cmd.ext 入侵此電子商務網站，進而值入後門程式等皆無所遁形，如 Get /cgi-bin/winnt/system32/cmd.exe?/c+dir。

在 Client 端發送一個 Request 封包到達網站(Server)後，網站就須回送於 Client 端一個 Response 封包(封包內容如圖 4)，只需查看回應封包的執行狀態碼，這資訊決定 Client 的請求是成功(「2」字作開始的狀態碼)、被導向至別處(「3」字作開始的狀態碼)、執行錯誤(「4」字作開始的狀態碼)或伺服器錯誤(「5」字作開始的狀態碼)，而本系統最主要判斷回應封包的狀態碼是否為 200，即表回送成功，而藉此封包內容可得知檔案大小(Content-Length)，型態(Content-Type)，如 image/jpg 等，及下載時間等相關資訊，由 Layer7 的分析方式，只要鎖定網頁名稱，就可得知特定網頁點擊率，再將蒐集到的資料加以匯整統計便能得知排行等相關實務資訊。




1. Browser 根據 URL 指示向 Server 發出 Query
2. Server 根據 http 請求傳送對應的檔案內容
3. Browser 顯示並解釋回送內容

**圖 2 HTTP 協定運作方式**

```
⊟ Hypertext Transfer Protocol
    GET /Top_Files/sethome.GIF HTTP/1.1\r\n
    Accept: */*\r\n
    Referer: http://www.mis.stut.edu.tw/Top.asp\r
    Accept-Language: zh-tw\r\n
    Accept-Encoding: gzip, deflate\r\n
    User-Agent: Mozilla/4.0 (compatible; MSIE 5.C
    Host: www.mis.stut.edu.tw\r\n
    Connection: Keep-Alive\r\n
```

**圖 3 Request packet format**

```
⊟ Hypertext Transfer Protocol
    HTTP/1.1 200 OK\r\n
    Server: Microsoft-IIS/5.0\r\n
    Date: Wed, 13 Nov 2002 13:39:08 GMT\r\n
    Content-Type: image/gif\r\n
    Accept-Ranges: bytes\r\n
    Last-Modified: Wed, 29 Aug 2001 07:19:00 G
    ETag: "0ca9edf5a30c11:94a"\r\n
    Content-Length: 164\r\n
```

**圖 4 Response packet format**

## 程式架構

此系統是採用 Linux 為主機，因此所利用的撰寫語言為 Linux 上的 C Language，而撰寫中所採用的幾個主要程式技巧如下：Sharing Memory、Link-List、Hash Table、Thread，而資料庫方面是使用 MySQL。

其程式架構流程如下圖（圖 5），首先我們利用之前所提到的全接收模式將所傳送過來的原始封包接收至 Buffer 中，再透過程式將原始資料讀取出來丟至 Filter 中進行過濾，產生我們所需要的資訊，此時我們不會把所得資訊儲存到資料庫，我們會先將此資訊儲存在我們用 Hash 與 Link-List 產生的記憶體結構中，以便更新接收回應封包，另外在程式剛開始時我們呼叫幾個 Threads 來做不同的工作，以達到多工的目的，也可以加增加程式的執行效率，其中一個 Thread 會負責做 Flags 的切換，主要是切換記憶體寫入的位置與從記憶體中讀出資料寫入檔案的動做，以便讓記憶體中所存的資料能夠儲存到資料庫中，而資料從記憶體中讀出寫到資料庫是透過另一個 Thread，在圖中（圖 5）負責將原始資料做過慮的 Filter 是屬於另一個 Thread，由圖中可以看記憶體資料是共享的，而每個 Thread 專門負責自己的工作，達到我們想增加執行效率的目的。

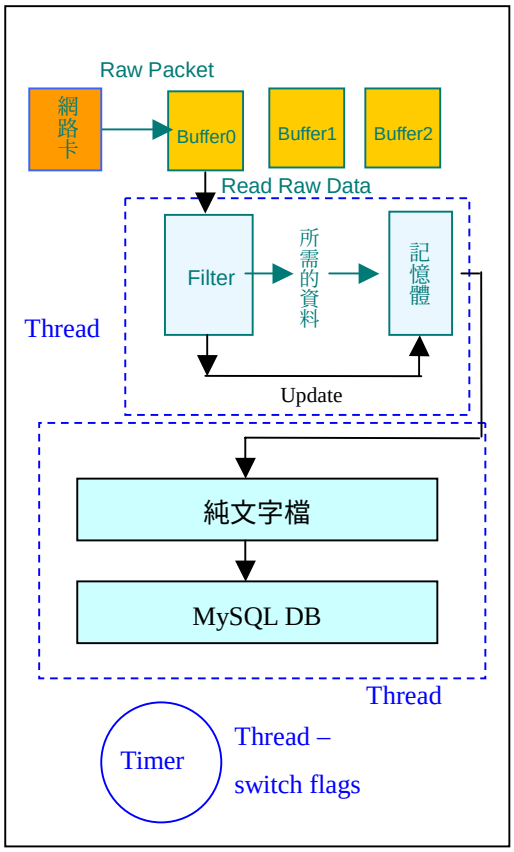



**圖 5 程式架構流程圖**

另外為了加快記憶體中更新資料的速度，我們透過 Hash 的技巧結合 Link-List 的結構來當做記憶體中儲存資料的結構，在 Hash 的技巧中，我們知道在找尋比對資料時的速度取決於 Hash Node 是否有平均分散，假如 Hash Table 中的 Node 越分散就代表需要找的次數越少，甚至可以直接命中找到，我們的做法是透過將取得的部份資訊互相執行 XOR 的動作，產生 Node 散佈平均的 Hash Table，記憶體中的結構如下圖（圖 6 記憶體結構）

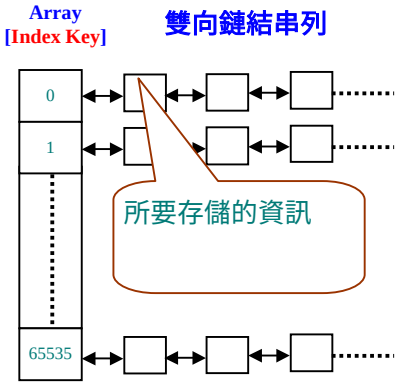



圖 6 記憶體結構

**使用者介面**：透過我們所設計的主系統處理後的資訊都會儲存在資料庫中，此時我們就可以透過 PHP 與 MySQL 的結合，設計出一個 Web 介面，將這些資訊呈現出來，藉由以上的資料經整理後便可輸出以下等各項結果：

1. 網站資源存取狀況分析
   - 由網頁的點擊率得知：熱門網頁排行、冷門網頁排行、進站網頁排行、離站網頁排行、單一網頁瀏覽排行，可由 Layer7 分析。
   - 由檔案或網頁的開啟數得知：最常存取目錄、最常下載的檔案、最常被索取檔案類型，可由 Layer7 分析。
2. 網站頻寬使用診斷
   - 檔案被下載的數量及檔案大小得知：每日頻寬使用量、各時段頻寬使用量、使用頻寬最多之檔案排行，需由 Layer3 分析。
   - IP 排行，需由 Layer3 分析。
   - Port 排行，需由 Layer4 分析。
3. 訪客來源及時段流量分析
   - 依 IP 來源得知：訪客人數、訪客來源、訪客所屬網域、訪客所屬單位，需由 Layer3、Layer7 分析。
   - 依網頁及檔案點擊數得知：每日觸擊數、每月點擊數、每人點擊數，需由 Layer3、Layer7 分析。
   - 依 IP 及時間得知：各時段的造訪人次，由 Layer3、Layer7 分析。

## 實驗測試

利用封包監聽技術不僅能得知訪客對網站的行為模式及相關的重要資訊，由於封包監聽是依序將訪客的瀏覽順序做記錄，所以更能真實重現訪客的瀏覽過程，本文針對 163.26.220.223 此 IP 為某一訪客做測試，連接至南台科計大學首頁、成功大學首頁，並將瀏覽過的網頁、圖檔及文件檔明細列出，並會在 Web 畫面中得知點擊數或下載次數（如圖 7），在圖中可以看見共有被點選過的 6 個網頁、下載過 74 個(jpg/gif)圖檔、及開過 1 個文件檔，若點選所列出的網頁連結，便能將使用者所瀏覽過的網頁重現，這些資訊都可以從資料庫中找出，假如使用者需要其他不同的剖析面去觀查，也可以自行透過 PHP 去連結資料庫產生自幾所要看的畫面，以上是針對應用層(Layer 7)分析，另外本文也秀出 IP 層(Layer 3)、傳輸層(Layer 4)的 IP 流量統計排行（如圖 8）及各種應用服務（如 WWW、FTP、Telnet、Emaile 等）Port 的使用量排行（如圖 9），並可依時段做明細查詢（如圖 10）。

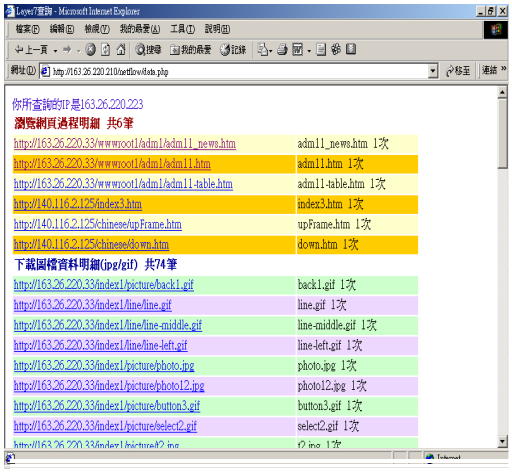



**圖 7 列出 163.26.220.223 瀏覽相關資料**

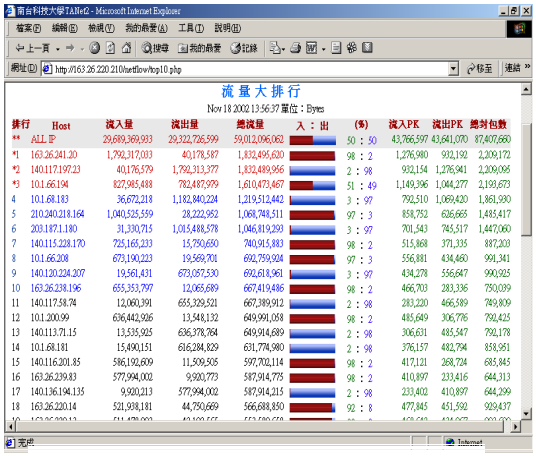



**圖 8 依流量做 IP 排行榜**

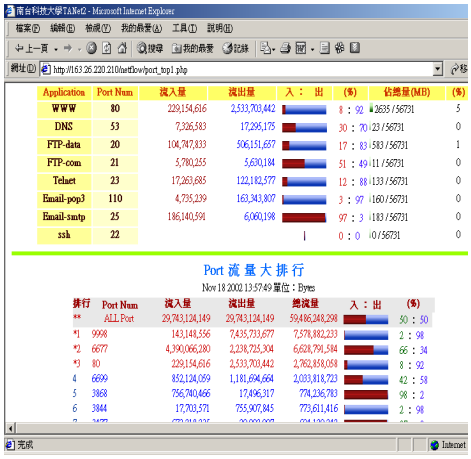



**圖 9 依各種應用服務流量做 Port 排行榜**

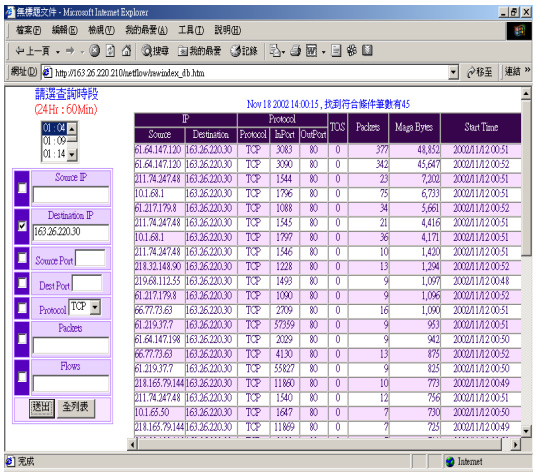



**圖 10 依時段查詢流量下載明細**

## 結論

目前網際網路的盛行，頻寬已不在是問題，WWW 的使用量與提供的資料服務越多，就越有衡量的需要與價值，透過封包監聽技術的統計分析，可以讓網站管理與資料提供人員，更加了解網路使用者的行為模式、與網站的服務成效，且將所得到的資料變成商機。

目前透過此研究成果已經達到當初所設定的目的，此研究成果也可以應用在許多方面，另外在研究的過程中發現，利用封包監聽方式可使得在使用者不被察覺的情行下，達到追蹤使用者從一開始進入網站瀏覽過哪些網頁、下載過哪些檔案等到離開網站為止，因此可任意的窺視別人在網路上的隱私權，所以應加強電子商務網站的網路安全，如將重要的資料傳送時使用加密的動作，這是一個值得討論及深思的問題。

關於封包監聽技術未來發展，可以著重於效能上，我們在研究發現，封包監聽技術效能的好壞，最主要的關鍵點在於網路頻寬的多寡，在現今高速的網路頻寬下，T1 已不敷使用，由其是大型電子商務的網站，由於需調配最佳效能，故需把主機做分散式處理，後端所架設的網站主機更是不下數十台，而對外出口的網路頻寬當然也就得跟著升級，在更速的 T3、Fast Ethernet，甚至是 Gigabit 的情況下，封包監聽技術的處理機制，對於封包遺失率、正確性等問題將面臨更大的挑戰，所以在未來需得做多方面的頻寬測試來改進整個效能以應付未來的趨勢。

封包監聽技術應用的範圍不只於電子商務，更能推廣至目前所盛行的校園網路教學，也是可利用同樣的處理機制來得知目前熱門或冷門的課程排行等相關訊息，在其它方面如果使用得當，相信封包監聽的技術日後也是相當重要的一門課題。

## 參考文獻

1. 陳伯榆，劉建男, “從 Unicode 入侵檢驗微軟 IIS 伺服器建置管理與防護”, 中正大學電信傳播所, TANet2001, I108

2. 賴源正，邱啟勝，簡士哲, “WebEyes 網頁監測系統”，國立台灣科技大學資訊管理所，TANet2001, W123
3. 連文雄，WWW 網站的管理與使用衡量的探討，國立中央大學圖書館通訊，第二十七期, 87 年 12 月，http://www.lib.ncu.edu.tw/c/book/n27/27-2.html
4. “Apache HTTP Server 日誌檔案”，http://itzone.info/dev/skyap/docs/logs.htm
5. “accesswatch”, http://accesswatch.com/
6. James Edwards , “Introducing HyperText Transfer Protocol (HTTP)”，http://linux.osso.org.co/documents/unix_unleashed_internet/ch21.htm
7. Kernel Korner: The Linux Socket Filter: Sniffing Bytes over the Network,” http://www.linuxjournal.com/article.php?sid=4659