Tools to Trace Path between Server and Web Applications over Internet

P Veeralakshmi
Department of Computer Science, Sourashtra College, Madurai-5 veeralakshmi@sourashtracollege.in

A Dhavapandiammal
Department of Computer Application, Sourashtra College, Madurai-5 davapandiammal@sourashtracollege.in

Abstract

The rise of the Internet and the worldwideweb had a wide revolution not only in network communications but also in application design and development. It is common to captured pieces of business functionality into distinct objects or components, and then made them available as self-contained Web services that can be accessed using Internet-based protocols and tools. Some tools Smartsniff, Httpnetworksniiffer, Winwebcrawler, Wireshark, Network Manager and WebLogAnalyzer are discussed to trace the path between client request and server response, history report of path travelled.

Keywords: network manager, protocols, report, webcrawler, wireshark

1. Introduction

The Internet is itself an enormous network, so whenever we access the Internet, we are using a network. The Web is based on a classical client-server architecture. Web servers that deliver static content only have to extract a requested document from the file system and pass it to the web browser protocol analyzer is basic tool for seeing the bits and bytes flowing through a network in human readable form. Network tools used to capture tcp/ip packets that pass through our network, and produce result. Every result message contains all necessary context information thus neither the server nor the client has to store context data.

2. Smartsniff

Smartsniff is a networking utility that allows you to capture tcp/ip packets that pass through your network adapter, and view the captured data as sequence of conversations in ascii mode(for text-based protocols, like http,smtp,post3 and ftp) or hex as dump.(for non-text base protocols, like dns).

for every request, the following information is displayed:

- protocol
- remote port
- remote host
- service name
- packets

3. Httpnetworksniffer

Httpnetworksniffer is a packet sniffer tool that captures all http requests/responses sent between the web browser and the web server and displays them in a simple table. for every http request, the following information is displayed: host name, http method (get, post, head), url path, user agent, response code, response string, content type, referer, content encoding, transfer encoding, server name, content length, cookie string, and more. you can easily select one or more http information lines, and then export them to text/html/xml/csv file or copy them to the clipboard and then paste them into excel.

4. web crawler

Figure 2: Httpnetworksniffer tool

Figure 1: captured protocols through smartsniff
A web crawler is an internet bot that systematically browses the world wide web, typically for the purpose of web indexing, web search engines and some other sites use web crawling or spidering software to update their web content or indexes of others sites' web content. Web crawlers can copy all the pages they visit for later processing by a search engine that indexes the downloaded pages so that users can search them much more quickly.

A web crawler starts with a list of urls to visit, called the seeds. As the crawler visits these urls, it identifies all the hyperlinks in the page and adds them to the list of urls to visit, called the crawl frontier. Urls from the frontier are recursively visited according to a set of policies. If the crawler is performing archiving of websites it copies and saves the information as it goes. Such archives are usually stored such that they can be viewed, read and navigated as they were on the live web, but are preserved as 'snapshots'.

The large volume implies that the crawler can only download a limited number of the web pages within a given time, so it needs to prioritize its downloads. The high rate of change implies that the pages might have already been updated or even deleted.

The number of possible urls crawled being generated by server-side software has also made it difficult for web crawlers to avoid retrieving duplicate content. Endless combinations of http get (url-based) parameters exist, of which only a small selection will actually return unique content. For example, a simple online photo gallery may offer three options to users, as specified through http get parameters in the url. If there exist four ways to sort images, three choices of thumbnail size, two file formats, and an option to disable user-provided content, then the same set of content can be accessed with 48 different urls, all of which may be linked on the site. This mathematical combination creates a problem for crawlers, as they must sort through endless combinations of relatively minor scripted changes in order to retrieve unique content.

The behavior of a web crawler is the outcome of a combination of policies: a selection policy that states which pages to download,

- a re-visit policy that states when to check for changes to the pages,
- a politeness policy that states how to avoid overloading web sites, and
- a parallelization policy that states how to coordinate distributed web crawlers.

In Fig 4, win web crawler’s running process captured. Win web crawler will query all popular search engines, extract all matching urls from search results, remove duplicate urls and finally visits those websites and extract data from there. Win web crawler send queries to search engines to get matching website urls. Next it visits those matching websites for data extraction. How many deep it spiders in the matching websites depends on depth setting of external site tab.

4.1 Depth
Here you need to tell win web crawler - how many levels to deep down within the specified website. If you want win web crawler to stay within first page, just select process first page only. A setting of "0" will process and look for data in whole website. A setting of "1" will process index or home page with associated files under root dir only. These 2 urls are same. When we set to ignore urls case, then win web crawler convert all urls to lowercase and can remove duplicate urls however - some servers are case-sensitive and we shouldn’t use this option on those special sites.

5. Wireshark
Wireshark is a free and open-source packet analyzer. It is used for network troubleshooting, analysis, software and communications protocol development. Wireshark allows the user to put network interface controllers that support promiscuous mode into that mode, in order to see all traffic visible on that interface, not just traffic addressed to one of the interface's configured addresses and broadcast/multicast traffic. However, when capturing with a packet analyzer in promiscuous mode on a port on a network switch, not all of the traffic travelling through the switch will necessarily be sent to the port on which the capture is being done, so capturing in promiscuous mode will not necessarily be sufficient to see all traffic on the 

special issue published in Int. Jnl. Of Advanced Networking and Applications (IJANA)

Page 63
network, port mirroring or various network taps extend capture to any point on the network.

Figure 5: Wireshark tool (packet analyzer)

in Fig 6 datalink layer is explained, once packet is captured in this layer it will show source and destination address.

Figure 6: Data link Layer in Wireshark

in Fig 7 network layer is responsible for defining an internet protocol and forward packets.

Figure 7: Network Layer in Wireshark

in Fig 9 can see the port details established.

Figure 9: Application Layer in Wireshark

Figure 8: Transport Layer in Wireshark

a transport layer provides end-to-end or host-to-host communication services for applications within a layered architecture of network components and protocols. in figure 8 upper packet is captured can see ip addresses in the source and destination columns.

6. Network manager:

network manager tool manages load balancing, routes and connections. In fig 10 tool local ip address and public ip address, also upload and download data rates per second.

Figure 10: Network Manager Tool
in Fig 11 multiple interfaces can be configured to check load balancing.

in Fig 12 routes destination, mask, gateway, type, protocol, metric datas are retrieved router metrics are metrics used by a router to make routing decisions. It is typically one of many fields in a routing table. Router metrics can contain any number of values that help the router determine the best route among multiple routes to a destination. Protocol type is used to define status of request page. Load balancing is used configure gateway and interface. It must be equal to load the routes in easy way.

in Fig 13 the connections captured are shown. It shows which process is requested and it displays local address, local port, remote address, remote port, which type of protocol and state is displayed.

7. web log history

here in Fig 14 we have captured web log history of gmail. In network it will display line by line. Each line describes name path, method, status: 200 ok, domain: mail.google.com, size content, type: text/javascript, time latency in minutes, timeline.

If we click on name path it will show all details:
- Remote address: 125.236.174:443
- Request: https://talkgadget.google.com/u/0/talkgadget/_/frame?v=1425318628&hl=en, request method: GET status code: 200 ok
- Request header: provisional headers are shown, accept text/html,application/xhtml+xml, application/xml; q=0.9, image/webp,*/*; q=0.8
- Referer: https://mail.google.com/mail/u/0/user-agent mozilla/5.0 (windows nt 6.1) applewebkit/537.36 (khtml, like gecko) chrome/40.0.2214.45 safari/537.36, response headers alternate-protocol: 443:quic,p=0.08, 443:quic,p=0.08, cache-control private, max-age=31536000, private, max-age=31536000, content-encoding: gzip, content-type: text/html, date: Wed, 14 May 2015 04:59:32 GMT expires: Wed, 04 Mar 2015 04:59:32 GMT, server: gse.

7.1 status code

- 4xx codes indicate a client side error (bad request, page not found, request timeout, forbidden, etc.)
- 5xx codes indicate a server side error (service unavailable, gateway timeout, internal server error - web server is busy, etc.).
- 1xxx codes indicate winsock error codes (host not found, no route found, host is down, etc).
- 1xx codes are informational - not used, but reserved for future use
- 3xx codes indicate that the request has been redirected and further action must be taken in order to complete the request
- 2xx codes indicate that the action was successfully received, understood, and accepted.

8. Conclusion:

As network servers have become capable of supporting Internet-based services, software developers have revised
new programming paradigms to take advantage of the widespread availability of these services. By using tools we can trace route for understanding where problems are in the Internet network and for getting a detailed sense of the Internet itself. Tools are very efficient and user friendly. For report generations about network packets, routers and protocols these tools are highly useable.

References:


