Análisis del comportamiento autosimilar con distribución Pareto del tráfico Ethernet

Santiago Pérez, Higinio Facchini, Alejandro Dantiacq, Gustavo Mercado, Luis Bisaro
{santiago.perez, higinio.facchini, alejandro.dantiacq, gustavo.mercado, luis.bisaro}@gridtics.frm.utn.edu.ar
GRID TICs
(Grupo de Investigación y Desarrollo en TICs)
Facultad Regional Mendoza,
Universidad Tecnológica Nacional
Mendoza Argentina
0261-5244576

Jesús Rubén Azor
Facultad de Ingeniería
Universidad de Mendoza
Mendoza Argentina

Resumen

El análisis de colas ha sido y es de enorme utilidad para los diseñadores de redes y analistas de tráfico, a efectos de planificar las capacidades de los elementos de red y predecir su rendimiento. Estos análisis dependen de la naturaleza Poisson del tráfico de datos. Sin embargo, muchos resultados predichos a partir del análisis de colas difieren significativamente del rendimiento observado en la realidad. Diversos estudios han demostrado que para algunos entornos el patrón de tráfico es autosimilar, en lugar de Poisson. Este concepto está relacionado con otros más conocidos como son los fractales y la teoría del caos. Desde principio de los años 90 se comenzaron a publicar documentos referidos a la autosimilitud del tráfico de Ethernet. La línea de investigación se desarrolla a través de los siguientes apartados: 1) Características de la autosimilitud, 2) Tráfico de datos autosimilar, 3) Tráfico de datos Ethernet, 4) Caso de Estudio experimental de tráfico Ethernet, 5) Análisis con distribución Pareto, y 6) Análisis de bondad de ajuste con la prueba de Kolmogorov-Smirnov.

Palabras claves: tráfico, autosimilitud, Ethernet, Pareto, bondad de ajuste

1 AUTOSIMILITUD

La autosimilitud es un concepto tan importante que, en cierta forma, resulta sorprendente que sólo se haya aplicado al análisis del tráfico de comunicaciones de datos en los últimos tiempos [1]. La autosimilitud se puso de manifiesto en una afirmación efectuada por Manfred Schroeder [2]. El concepto unificador que subyace a los fractales, al caos y a las leyes de potencia es la autosimilitud. La autosimilitud, o invariancia frente a campos de escala o tamaño, es un atributo de muchas leyes de la naturaleza.
Los fenómenos autosimilares tienen el mismo aspecto o comportamiento cuando se visualizan con distintos grados de ampliación o a distintas escalas en una cierta dimensión. La dimensión puede ser el espacio (longitud, anchura) o el tiempo. En ese trabajo interesan las series temporales y los procesos estocásticos que muestran una autosimilitud con respecto al tiempo.
Es posible que la característica que más se destaque, desde el punto de vista del rendimiento de la red, es la persistencia de los agrupamientos. En un tráfico Poisson, los agrupamientos se producen a corto plazo (en una escala temporal pequeña), pero se van suavizando a largo plazo. Esto da lugar a la observación de que el análisis de colas tradicional, que supone un tráfico de Poisson, no prediga con precisión el rendimiento de un tráfico autosimilar.

2 TRÁFICO DE DATOS AUTOSIMILAR

Una señal periódica determinista está caracterizada por su invariancia respecto a la traslación temporal: la señal es idéntica si se traslada en el tiempo por múltiplos del periodo. Por el contrario, para un proceso estocástico estacionario, los estadísticos del proceso son invariantes frente a traslaciones temporales: la media y la varianza no dependen del tiempo y la función de autocorrelación depende sólo de una diferencia temporal.
Las señales autosimilares deterministas son invariantes frente a cambios de escala. Para un proceso estocástico, se puede decir que los estadísticos del proceso no cambian cuando cambia la escala temporal.
Tanto desde el punto de vista cualitativo como desde el cuantitativo, el proceso carece de una escala característica: el comportamiento medio del proceso a corto plazo es igual a su comportamiento medio a largo plazo. Esta autosimilitud no determinista es bastante frecuente, tanto en los fenómenos naturales como en los humanos; se puede apreciar en los paisajes naturales, en la distribución de los terremotos, en las olas de los océanos, en el flujo turbulento, en las fluctuaciones de la Bolsa y en el patrón de errores y en el tráfico de datos en los canales de comunicaciones.

Parámetro de Hurst

El parámetro H, que se denomina parámetro de Hurst, o parámetro de autosimilitud, es una medida clave de la autosimilitud. Más exactamente, H es una medida de la persistencia de un fenómeno estadístico y es un indicador de la longitud de la dependencia a largo plazo de un proceso estocástico. Un valor de H=0,5 indica la ausencia de dependencia a largo plazo. Cuanto más próximo esté H a 1, mayor será el grado de persistencia o de dependencia a largo plazo.

Distribución de Pareto

Se pueden definir procesos estocásticos autosimilares con distribuciones de cola hiperbólica. Uno de los atractivos de las distribuciones hiperbólicas es que dan lugar a modelos de simulación manejables.
Las distribuciones de cola periódica sirven para caracterizar densidades de probabilidad que describen procesos de tráfico como los tiempos entre llegadas de paquetes y las longitudes de

ráfaga. En general, una variable aleatoria con distribución de cola hiperbólica muestra una varianza infinita y posiblemente una media infinita. Las variables aleatorias con distribución de cola hiperbólcia contienen valores muy grandes con una probabilidad no despreciable. Generalemente, si se muestra una de estas variables, el resultado incluirá muchos valores relativamente pequeños, pero también unos pocos valores relativamente grandes.
La distribución de cola hiperbólica más sencilla es la distribución de Pareto con parámetros a y b, cuyas funciones de densidad y de distribución son

$$f(x) = \frac{ab^a}{x^{a+1}}$$

$$F(x) = 1 - \left(\frac{b}{x}\right)^a$$

para valores de $x \geq b$

El parámetro ***b*** especifica el valor mínimo que puede tomar la variable aleatoria. El parámetro ***a*** determina la media y la varianza de la variable aleatoria. Cuando se comparan las funciones de densidad de Pareto y exponencial en una escala semilogarítmica, se observa que en esta escala la función densidad exponencial es una recta, reflejando el decrecimiento exponencial de la distribución. El final de la distribución de Pareto decrece mucho más lentamente que la exponencial; de aquí viene el nombre de cola hiperbólica.
La distribución de Pareto se ha observado en una amplia gama de fenómenos procedentes de las ciencias sociales y físicas, y del mundo de las comunicaciones [3]. La cola hiperbólica de ciertas variables de las redes (por ejemplo, los tamaños de los archivos y las duraciones de las conexiones) son la causa seminal subyacente de la dependencia de largo alcance y de la autosimilitud que se observa en el tráfico de red [4].

3 TRÁFICO DE DATOS ETHERNET

El artículo fundamental del estudio de los datos de tráfico autosimilar es «On the Self-Similar Nature of Ethernet Traffic» (La naturaleza autosimilar del tráfico de Ethernet), que posteriormente sería corregido y aumentado en [5]. Este documento contradijo la idea de que un simple análisis de colas basado en la suposición de que el tráfico fuera de Poisson pudiera modelar adecuadamente todo tráfico de red. Empleando una masiva cantidad de datos y un cuidadoso análisis estadístico, el artículo manifiesta que, para el tráfico de Ethernet, se requiere un nuevo planteamiento de modelado y de analísis.
Por ello, en las simulaciones se prefiere modelar los periodos de tiempo de tráfico, con distribuciones de varianza infinita, utilizando en particular la distribución de Pareto. Esto da como resultado una distribuciçon de elevada varianza, con muchas ráfagas muy cortas, muchas ráfagas largas y algunas ráfagas muy largas. Esto ha permitido determinar el origen de las descrepancias, por ejemplo, entre el tiempo real de espera y el tiempo estimado de espera obtenidos mediante el uso de la teorçia de colas convencional usando Poisson.

4 CASO DE ESTUDIO EXPERIMENTAL DE TRÁFICO ETHERNET

Los métodos de colección de tramas de red Ethernet, son el punto de partida para el entendimiento del comportamiento del tráfico y de los nodos de red [6]. Pueden ser clasificadas en tres categorías: 1) métodos basados en polling, los cuales registran las asociaciones de los nodos de red en intervalos de tiempo periódicos, usando el protocolo snmp, o algún software de tracking de asociación sobre los nodos, 2) métodos basados en programas que registran eventos online/offline de usuarios de red usando un servidor de logeo (syslog) como archivos de sesión, de dhcp, de traps, etc. y 3) métodos basados en programas sniffers que colectan el tráfico de la red en la medida que se produce.

A los fines del trabajo, en el que se pretende analizar algún patrón característico del tráfico de red Ethernet, es importante incluir en la colección de trazas, las tramas de tráfico con la mayor cantidad de información posible usando sniffers. Existe una extensa variedad de estos programas para analizar tramas y trafico de red, como TCPdump [7], IPtraf [8], Wireshark [9] (ex Ethereal), NTOP (Network TOP) [10], entre otros.

Para el estudio experimental, se tomará una muestra, utilizando el sniffer EherPeek [11]. Una vez colectadas las tramas, se comenzará su análisis en la misma herramienta, dado que posee una gran flexibilidad en su interfaz, identificando en cada una el protocolo de capa superior TP/IP involucrado, su longitud en bytes, etc y especialmente el instante de tiempo de muestreo. Además, da la posibilidad de exportar los datos a una planilla de cálculo para facilitar su manipulación y representación.

5 ANÁLISIS CON DISTRIBUCIÓN PARETO

En estadística, la distribución Pareto, formulada por el sociólogo Vilfredo Pareto, es una distribución de probabilidad continua con dos parámetros *a* y *b* cuya función de densidad para valores $x \geq b$ es:

$$f(x) = \frac{ab^a}{x^{a+1}}$$

Y su función de distribución es:

$$F(x) = 1 - \left(\frac{b}{x}\right)^a$$

El valor esperado y la varianza de una variable aleatoria X de distribución Pareto son

$$E[X] = \frac{ab}{a-1}$$

$$V[X] = \frac{ab^2}{(a-1)^2(a-2)}$$

La distribución de Pareto, puede expresarse como una función f(x,a,b), de la siguiente forma:

$$f(x,a,b) := \frac{a \cdot b^a}{x^{a+1}}$$

Asignando a los parámetros los valores a=0,9 y b=1, se puede generar el vector E (vector de valores esperados) para la distribución de Pareto, con x variando entre 1 y 10.

$$i := 1..10$$

$$E_{i-1} := 1 \cdot \int_{i}^{i+1} f(x, 0.9, 1)\, dx$$

6 ANÁLISIS DE BONDAD DE AJUSTE CON LA PRUEBA DE KOLMOGOROV-SMIRNOV.

El uso de la Estadística es de gran importancia en la investigación científica. Casi todas las investigaciones aplicadas requieren algún tipo de análisis estadístico para que sea posible evaluar sus resultados. Por ejemplo, los tests o dócimas paramétricos y no paramétricos.
Dentro de las pruebas no paramétricas, se destacan las pruebas de Kolmogorov-Smirnov para una y dos muestras. Se han propuesto diferentes métricas para describir y comparar utilizando diferencias entre distribuciones acumuladas. La prueba unimuestral de Kolmogorov-Smirnov es una prueba de Bondad de Ajuste apropiada para este caso en que se está usando la distribución Pareto [13]. Es más eficiente que la prueba χ^2 en muestras pequeñas, y no se aplica a distribuciones discretas.
La prueba unimuestral se funda en la diferencia absoluta máxima D entre los valores de la distribución acumulada de una muestra aleatoria de tamaño n, y una distribución teórica determinada. Para decidir si esta diferencia es mayor de la razonablemente esperada con un nivel de significación α, se buscan los valores críticos de D en Tablas apropiadas.

7 CONCLUSION

En este documento, se han relacionado los temas de autosimilitud, con el tráfico Ethernet, la distribución de Pareto y la prueba de Kolmogorov-Smirnov.
El volumen de los trabajos y literatura sobre tráfico de datos es creciente, y el tema de la autosimilitud ha significado el principio de un nuevo examen del rendimiento del tráfico de datos, las técnicas de modelado, y control de tráfico, entre otros.
En la investigación se pretende verificar a través de un estudio experimental y usando la prueba de Kolmogorov-Smirnov, que el tráfico de datos Ethernet responde efectivamente a la distribución Pareto.

8 REFERENCIAS

[1] Stalling, W.; "Redes e Internet de Alta Velocidad. Rendimiento y Calidad de Servicio", Pearson, 2003.

[2] Schroeder, M.; "Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise", Nueva York, Freemam, 1991.

[3] Paxson,V. y Floyd, S.; "Wide Area Traffic: The Failure of Poisson Modeling", IEEE ACM Transacctions on Networking, 2000.

[4] Park, K. y Williams, W.; "Self-Similar Network Traffic and Performance Evaluation", Nueva York, Wiley, 2000.

[5] Leland, W; Taqqu, M; Willinger, W. y Wilson, D; "On the Self-Similar Nature of Etehrnet Traffic", IEEE/ACM Transaccions on Networking, Febrero 1994

[6] Perez, S; Mercado, G. y Facchini, H.; "Análisis y Determinación de Patrones de Tráfico de Protocolos en redes LAN", WICC 2007, 2007.

[7] http://www.tcpdump.org

[8] http://iptraf.seul.org

[9] http://www.wireshark.org

[10] http://www.ntop.org/

[11] http://ether-peek.softonic.com/mac

[12] http://www.ptc.com/products/mathcad/

[13] http://www.um.edu.ar/math/estadis/programa.htm