Científica Vol. 11 Núm. 1 pp.27-32
© 2007 ESIME-IPN. *ISSN 1665-0654*. Impreso en México

Análisis fractal en la predicción del tráfico en el servidor de correos electrónicos del Instituto Politécnico Nacional

Miguel Patiño-O.
Iván Campos-S.

Sección de Estudios de Posgrado e Investigación,
Escuela Superior de Ingeniería Mecánica y Eléctrica,
Instituto Politécnico Nacional. Unidad Profesional "Adolfo López Mateos". Col. Lindavista, México, DF, 07738.
MÉXICO.

email: icampos@ipn.mx

Recibido el 25 de mayo de 2006; aceptado el 31 de octubre de 2006.

1. Resumen

En el presente trabajo se estudia la dinámica del servidor de correos electrónicos del Instituto Politécnico Nacional (IPN) a través de la implementación de técnicas estadísticas y geometría fractal. La distribución estadística que mejor ajusta los datos del tráfico y de las diferencias del tráfico del servidor de correos electrónicos del IPN es Log-Logistic que presenta un comportamiento de ley de potencia. Las distribuciones encontradas en las series de datos de las diferencias del tráfico del servidor de correos son simétricas. A través del análisis fractal, se obtuvieron los exponentes de escalamiento del servidor de correos electrónicos que fueron calculados por cinco métodos de trazado autoafín. Los resultados obtenidos demuestran antipersistencia en el sistema, y un comportamiento de invarianza de escala en el tráfico y en las diferencias del tráfico del servidor de correos electrónicos del Instituto Politécnico Nacional.

Palabras clave: análisis estadístico, análisis fractal, exponente de Hurst, ley de potencia, sistemas complejos.

2. Abstract (Fractal Analysis on the Traffic Prediction in the Electronic Mails Server of the Instituto Politecnico Nacional)

In the present work the dynamics of the National Polytechnic Institute Electronic Mails Server (NPIEMS) is studied, through the implementation of statistical techniques and fractal geometry. The statistical distribution that better fits the traffic and the traffic differences data of the NPIEMS is Log-Logistic, whose presents a power law behavior. The distributions found in the series of the traffic differences data of the mails server are symmetrical. Through the fractal analysis, the scale exponents of the electronic mails server were obtained, calculated by five self-affine methods. The results demonstrate antipersistence in the system, and a scale invariance behavior in the traffic and the traffic differences of National Polytechnic Institute Electronic Mails Server.

Key words: statistical analysis, fractal geometry, Hurst exponent, power law, complex systems.

3. Introducción

En la actualidad el uso de la red Internet ha crecido aceleradamente, de tal forma que muchas de nuestras actividades, inclusive las más cotidianas, están relacionadas con ella. El aumento en el número de computadoras y programas que se conectan a la red ha generado un considerable incremento en el tráfico de la información.

Empresas, instituciones y usuarios de Internet, pronostican la dinámica que tendrá la red, debido a que es esencial para su rendimiento. El conocimiento del comportamiento de las redes permite mejorar la planeación estratégica de las *tecnologías de la información* (TI), específicamente en las decisiones de inversión relacionadas con éstas.

Los estudios de los datos del tráfico de las redes de comunicaciones, se han hecho con medidas en las diferentes capas del modelo de referencia [1, 2, 3, 4].

La red Internet es vista y considerada como un sistema complejo [5, 6, 7]. Una de las áreas que estudia a los sistemas complejos es la física estadística, y esto lo hace a partir de la dinámica que éstos manifiestan y no a través de la descripción de sus componentes. Esto permite abordar, de una forma conceptualmente unificada, una variedad de sistemas con diversos componentes, pero con comportamientos globales

similares [2]. Por lo tanto, pueden aplicarse conceptos de la física estadística al análisis del comportamiento global de la red Internet.

4. Desarrollo

4.1 Caracterización y modelación de la dinámica del servidor de correos electrónicos

Para la modelación y caracterización estadística del tráfico del servidor de correos electrónicos del IPN se llevan a cabo un conjunto de actividades que están basadas en una metodología de trabajo que se compone de tres enfoques complementarios: investigación, desarrollo y diagnóstico y evaluación de resultados [8, 9, 10]; además se emplean conceptos de la física estadística, probabilidad y geometría fractal. La figura 1 muestra el conjunto de actividades llevadas a cabo para la modelación y caracterización estadística del tráfico del servidor de correos.

4.1 Análisis estadístico

Para llevar a cabo el análisis estadístico se obtienen, transforman y filtran los datos de los archivos de registros del servidor de correos del IPN (correos electrónicos de los LOG de transacciones). Por otro lado, estos archivos se guardan en una base de datos para generar, caracterizar y analizar series temporales a diferentes escalas de tiempo (un segundo, 10 segundos, 20 segundos, 30 segundos, 40 segundos, 50 segundos, un minuto, una hora).

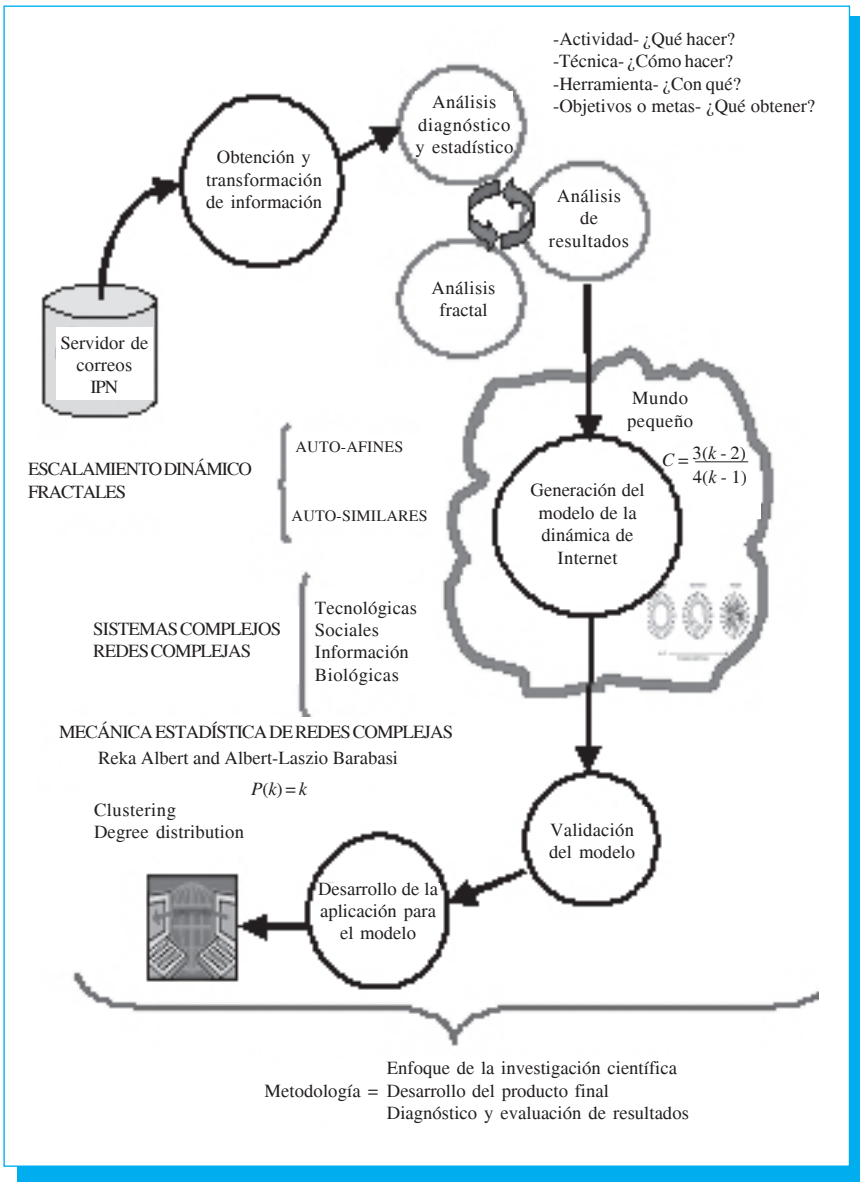

Fig. 1. Diagrama de procesos para la modelación y caracterización estadística del tráfico del servidor de correos electrónicos del IPN.

En el análisis estadístico, y empleando el software @Risk 4.5 [11] y Excel, se determinó que la distribución estadística que mejor ajusta el comportamiento histórico de las series de tiempo del tráfico del servidor de correos electrónicos y de las diferencias del tráfico, es la distribución Log-Logistic para todas las series.

La distribución Log-Logistic es una distribución de cola pesada (comportamiento de ley de potencia). La característica relevante de las leyes de potencia es que éstas son funciones cuya forma es de invarianza de escala, cuya simetría define la geometría fractal, por ende, la mayoría de las propiedades de los fractales están expresadas mediante leyes de potencia [12, 13, 14, 15, 16].

Es importante señalar que el software @ Risk 4.5 fue desarrollado para analizar situaciones sensibles al riesgo, ordena las distribuciones estadísticas empezando con las que mejor ajustan los datos, mediante tres criterios estadísticos: Chi-cuadrada, Anderson-Darling, y Kolmogorov-Smirnov [11].

Para la distribución Log-Logistic tenemos que:

Función de densidad: $$f(x) = \frac{\alpha\, t^{\alpha-1}}{\beta\left(1+t^{\alpha}\right)^{2}} \qquad (1)$$

Función acumulativa: $$F(x) = \frac{1}{1+\left(\frac{1}{t}\right)^{\alpha}} \qquad (2)$$

con $t \equiv \frac{x-\gamma}{\beta}$

donde:

α	Parámetro de forma continua.
β	Parámetro de escala continua.
γ	Parámetro de localización continua.
t	tiempo.

Las figuras 2(a) y 2(b) presentan las distribuciones Log-Logistic encontradas para el tráfico y para las diferencias del tráfico del servidor de correos electrónicos, en series de 10 segundos para el mes de octubre de 2005. Así mismo, la tabla 1 muestra los promedios de los parámetros estadísticos de las distribuciones para los meses de agosto de 2005 a enero de 2006.

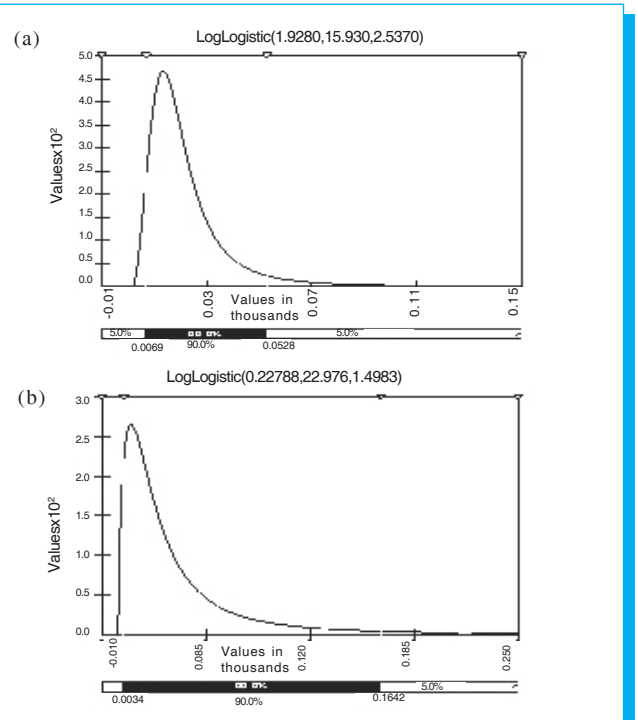

Fig. 2. (a) Distribución Log-Logistic para el tráfico del servidor de correos (b) Distribución Log-Logistic para las diferencias del tráfico del servidor de correos electronicos del IPN.

Tabla 1. Parámetros estadísticos obtenidos para las distribuciones Log-Logistic.

	Tráfico			Diferencias de tráfico		
	α	β	γ	α	β	γ
10 s	**1.69**	19.32	−0.9403	**1.44**	9.44	0.51
20 s	**2.15**	37.73	−2.0931	**2.52**	14.49	0.26
30 s	**2.35**	56.10	−2.2800	**2.55**	19.21	0.04
40 s	**2.41**	77.61	−1.8844	**2.56**	22.90	−0.08
50 s	**2.37**	108.67	−2.2059	**2.57**	26.37	−0.20
1 min	**2.40**	131.59	−1.1435	**2.57**	28.98	−0.28
Promedio	**2.23**	71.83	−1.7579	**1.53**	20.23	0.04
D.S.	**0.28**	42.73	0.5740	**0.05**	7.37	0.30
Máx.	**2.41**	131.59	−0.9403	**1.57**	28.98	0.51
Mín.	**1.70**	19.32	−2.2800	**1.44**	9.44	−0.28

De la tabla 1 el parámetro estadístico de localización continua α se establece en un valor de 1.53, que implica una distribución continua. Así mismo, las distribuciones encontradas en las series de datos de las diferencias del tráfico del servidor de correos son simétricas, lo que indica que la probabilidad de tener un incremento positivo o negativo es del 50%, es decir, mientras que en un periodo de tiempo el número de correos se incrementa, en el periodo siguiente disminuye, en el siguiente aumenta y así sucesivamente.

Para conocer la correlación entre los valores del tráfico y detectar la existencia de memoria en las series de tiempo se emplea la función de autocorrelación $C(n)$:

$$C(n) \propto \left(\left| n+1 \right|^{2H} - 2n^{2H} + \left| n-1 \right|^{2H} \right) \qquad (4)$$

donde $0 \le H \le 1$ es el exponente de Hurst. Por lo tanto, la función de autocorrelación de series autoafines despliega un comportamiento de escalamiento de la forma:

$$C(n) \propto 1 - n^{2H} \qquad (\text{si } n \to 0) \qquad (5)$$

y

$$C(n) \propto n^{-2(1-H)} \qquad (\text{si } n \to \infty) \qquad (6)$$

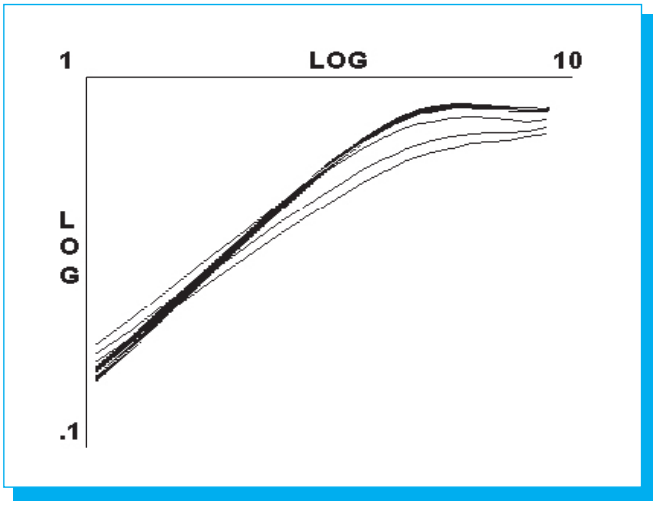

Fig. 3. Gráficas de la autocorrelación para series con periodos de 10 segundos, para los meses agosto 2005 - enero 2006.

Los resultados de la ecuación (5) y de los gráficos de la autocorrelación para series con periodos de 10 segundos, durante los meses de agosto de 2005 hasta enero de 2006 son mostrados en la figura 3 y la tabla 2.

De la tabla 2 se obtiene un valor promedio de $H = 0.2509$, lo cual indica antipersistencia en el sistema, es decir, si la tendencia del tráfico del servidor de correos electrónicos en un periodo es creciente, es probable que en el siguiente sea decreciente.

4.3 Análisis fractal

En este análisis todas y cada una de las series tiempo analizadas se dividieron en subseries con diferentes horizontes de tiempo (segundos, minutos, horas, días y meses) y se calculó el valor de los exponentes locales de escalamiento H (exponente de Hurst) para determinar la existencia de correlaciones en las series de tiempo. Se utilizaron cinco métodos autoafines con ayuda del software Benoit 1.2 [17]:

Tabla 2. Ecuaciones de línea de tendencia para la autocorrelación, para los mese agosto 2005 - enero 2006.

Mes	Ecuación	
08	$y = 0.5448x^{0.5547}$	$R^2 = 0.9123$
09	$y = 0.5558x^{0.5339}$	$R^2 = 0.9111$
10	$y = 0.5425x^{0.5143}$	$R^2 = 0.9469$
11	$y = 0.5856x^{0.4771}$	$R^2 = 0.9306$
12	$y = 0.5831x^{0.4569}$	$R^2 = 0.9118$
01	$y = 0.5662x^{0.4584}$	$R^2 = 0.9421$
02	$y = 0.5565x^{0.4577}$	$R^2 = 0.9542$

- Rango reescalado: $(R/S \propto \tau^{H})$ (7)
- Espectro de potencia: $(P \propto \tau^{-2H-1})$ (8)
- Rugosidad-longitud: $(SD \propto \tau^{H})$ (9)
- Variograma: $(V \propto \tau^{2H})$ (10)
- Ondoletas: $(W[X](a) = \langle |W(a,b)| \rangle_b \propto a^{H+1/2})$ (11)

El exponente de Hurst indica si las series de tiempo analizadas tienen comportamiento aleatorio ($H = 0.5$), persistente ($0.5 < H < 1$) o antipersistente ($0 < H < 0.5$). Otra forma de interpretar el exponente de Hurst, es a través de una medida de la variación de la información del tráfico en una serie de tiempo. Mientras el exponente de Hurst tiende a 1 las diferencias en el tráfico tienden a 0 y a su vez la dimensión fractal tiende a un valor de 1.

En el análisis se encontró un valor de H equivalente a 0.2519 para el tráfico, mientras que para las diferencias de tráfico H fue de 0.2537, lo que indica un comportamiento de antipersistencia. La tabla 3 presenta los resultados del promedio de los exponentes de Hurst por los métodos: rango-reescalado, variograma, rugosidad longitudinal, espectro de potencia y ondoletas. Por otro lado, cabe resaltar que el software Benoit 1.2 está limitado en algunos de sus métodos para el manejo de grandes volúmenes de datos, razón por lo cual se generó un programa en lenguaje "C" para el cálculo de los exponentes de Hurst. Los métodos de Benoit 1.2 que dieron un valor similar del exponente de escalamiento, junto con el programa en "C", fueron: rango reescalado (R/S analysis), rugosidad-longitud (*roughness-length*) y variograma (variogram). Los métodos de: espectro de potencia (power spectrum) y ondoletas (wavelets) muestran resultados con mucha variación y menos exactos, debido a errores estadísticos en los métodos.

De acuerdo a los valores mostrados en la tabla 3, el promedio del exponente de Hurst es menor a ½, lo que indica antipersistencia en el tráfico. Asimismo, la dimensión fractal es equivalente a $D = 2 - H$, obteniendo una dimensión de 1.7481 ± 0.0028 para el tráfico y de 1.7463 ± 0.0026 para las diferencias del tráfico del servidor, respectivamente.

Tabla 3. Resultados del promedio de los exponentes de escalamiento (exponentes de Hurst) por los métodos: rango reescalado, rugosidad-longitud, variograma, espectro de potencia y ondoletas.

		10 s	20 s	30 s	40 s	50 s	1 min	Prom.
Tráfico	Prom.	0.2369	0.2481	0.2511	0.2515	0.2603	0.2632	**0.2519**
	S.D.	0.0041	0.0027	0.0029	0.0026	0.0022	0.0025	**0.0028**
Diferencias de tráfico	Prom.	0.2549	0.2528	0.2563	0.2549	0.2542	0.2488	**0.2537**
	S.D.	0.0037	0.0028	0.0025	0.0022	0.0022	0.0022	**0.0026**

Por otro lado, se determinó que las series temporales de los datos del tráfico del servidor de correos y las diferencias del tráfico del servidor de correos muestran invarianza de escala, confirmando su comportamiento fractal.

La variación del tráfico está dada por:

$$\Delta t = \left| P(t) - P(t-\tau) \right| \propto T^{H} \qquad (12)$$

donde la varianza escala a una función potencial T^H donde $H = 0.2537$.

4.4 Predicción del tráfico

El conjunto de procedimientos para predecir el tráfico en el servidor de correos electrónicos mostrado en la figura 4, están basados en los resultados del análisis estadístico y análisis fractal. Para predecir la dinámica del tráfico del servidor de correos, se calculó el promedio del tráfico en las diferentes escalas de tiempo y el promedio del rango en las diferencias del tráfico del servidor de correos con ayuda del generador de trazas del software Benoit 1.2. Se tomó como base el promedio del rango de las diferencias del tráfico del servidor de correos, el promedio del coeficiente de escalamiento H= 0.2537 y el promedio del rango en las diferencias del tráfico (valor máximo menos valor mínimo del tráfico) igual a 910, generando 300 escenarios probabilísticos. Asimismo, se calcularon, para cada una de las 300 trazas, los 1 555 200 promedio del tráfico de Internet, proyectados para cada día de cada mes, para seis meses, a partir de los 1 555 200 datos de las diferencias de tráfico con sus signos positivos y negativos establecidos anteriormente.

Cada una de las 300 trazas se generaron de manera aleatoria, por lo que los 1 555 200 datos de cada traza son variables idéntica e independientemente distribuidas. Así mismo, se aplicó la propuesta de Sornette y Andersen[18] a cada una de las 300 trazas generadas. Al primer dato de cada traza se le asignó el signo positivo, al segundo dato el signo negativo, y así sucesivamente hasta que al dato 1 555 200 se le asignó el signo contrario al del dato 1 555 199.

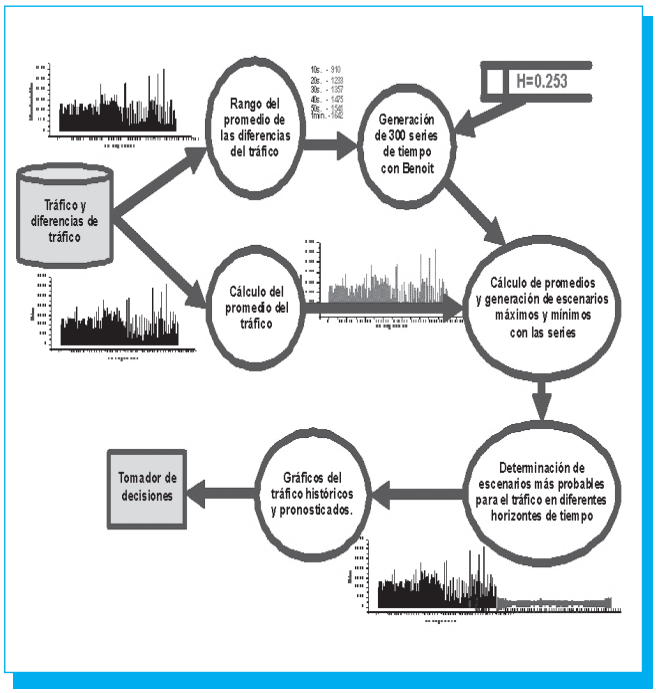

Fig. 4. Diagrama de procesos para la predicción del tráfico del servidor de correos electrónicos del IPN.

La fórmula para determinar el tráfico fue:

$$P(t) = P(t-\tau) + \Delta(\tau) \qquad (13)$$

El primer valor de $P(t-\tau)$ que se aplicó para calcular los 1 555 200 $P(t)$, fue el promedio de los últimos seis meses de agosto de 2005 a enero de 2006 del tráfico, para las series de 10 segundos. Posteriormente se calculó el valor máximo global y el valor mínimo global para cada punto, los cuales representan los valores más altos y más bajos del tráfico en la red. Así mismo se obtuvo el valor promedio de cada uno de los 1 555 200 datos de las 300 trazas que son los valores más probables para el tráfico del servidor de correos electrónicos del IPN.

La figura 5 muestra los valores pronosticados para el tráfico del servidor de correos electrónicos en los próximos seis meses, de abril a septiembre de 2006. El pronóstico se encuentra establecido para el servidor de correos electrónicos del IPN, sin embargo, la metodología puede ser aplicada en cualquier otro servidor o nodo de la red, esperando que los resultados de esto sean similares.

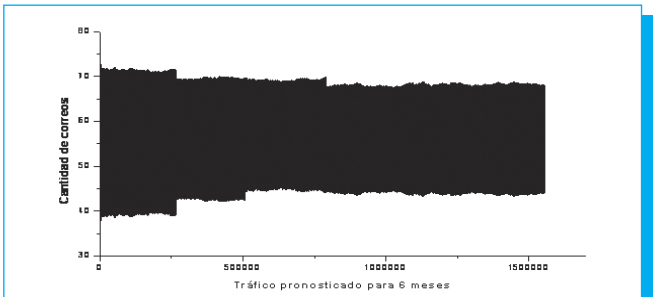

Fig. 5. Comportamiento esperado del tráfico del servidor de correos para un periodo de seis meses en intervalos de 10 s.

5. Conclusiones

Se desarrolló una metodología que caracteriza y pronostica el tráfico del servidor de correos electrónicos del IPN con un enfoque sistémico y sistemático, empleando técnicas y herramientas de la mecánica estadística. La metodología puede ser utilizada en cualquier otro servidor o nodo de la red dentro del Instituto Politécnico Nacional.

El análisis estadístico establece que la distribución probabilística Log-Logistic es la que mejor ajusta los datos del tráfico de la red. Los parámetros estadísticos de la distribución determinan que el parámetro de localización continua a es igual a 1.53, lo cual implica una distribución continua. Así mismo, las distribuciones encontradas en las series de datos de las diferencias del tráfico del servidor de correos son simétricas, lo que indica que la probabilidad de tener un incremento positivo o negativo es del 50%.

Se ha determinado que el valor de escalamiento *H* es de 0.2519 para el tráfico, mientras que para las diferencias de tráfico fue de 0.2537, lo que indica un comportamiento de antipersistencia del servidor de correos electrónicos del Instituto Politécnico Nacional.

6. Referencias

[1] Wetteroth Debbra. *OSI Reference Model for Telecomunications*. Mc Graw Hill. 2002. USA.

[2] K. B. Chong & K.Y. Choo. "Fractal Analysis on internet traffic time series". University of Fribourg. Econophysics Forum. arXiv.physics/0206012 v3. 2003. SWITZERLAND.

[3] Lent, Ricardo & Yamakawa, Peter. "Naturaleza Fractal del tráfico Internet". *TECNIA*, Vol. 8 No. 01, pp. 39-44, 1998. Lima - Perú.

[4] Vicari, Norbert. "Measurement and Modeling of WWW-Sessions". University of Wûrzburg - Institute of Computer Science - Research Report Series. Rep. No. 184. Sep. 1997 Germany.

[5] Park, Kihong. "The Internet as a Complex System". Deparment of Computer Sciences Purdue University.

[6] Newman, M. E. J. "The Structure and Function of Complex Networks". SIAM Review vol. 45, No. 2, pp. 167-256. 2003. USA.

[7] A. S. Balankin & Jesús Márquez González. "Fractal Behavior of Complex Systems". *Científica*, vol. 7 Núm. 3 pp 109-128.

[8] Checkland, Peter & Scholes, Jim; *Soft Systems Methodology in Action*; John Wiley; England 1990.

[9] John P. Van Gigch. *Teoría general de sistemas*, Editorial Trillas, México 1993.

[10] Hernández-Sampieri, R., Fernández-Collado C., Baptista, Lucio P. *Metodología de la Investigación*; McGrawHill, 1991.

[11] RISK@ http://www.palisade.com.

[12] B.B. Mandelbrot. The Fractal Geometry of Nature; W.H. Freeman, Nueva York, 1982.

[13] A-L Barabási & H.E. Stanley. *Fractal concepts in surface growth*; Cambridge University Press, Cambridge, 1995.

[14] Oswaldo Morales Matamoros. "Modelos Mecánicos de la Dinámica Fractal del Mercado Petrolero". Tesis Doctoral. SEPI- ESIME-Zacatenco. IPN. 2004.

[15] T. Vicsek. *Fractal growth phenomena*; World Scientific, 1989.

[16] B.B. Mandelbrot. *Fractals in Physics*; Holland, Amsterdam, 1986, pp. 3.

[17] BENOIT 1.3 http://www.scioncorp.com.

[18] D. Sornette & Andersen, "Increments of Uncorrelated Time Series Can Be Predicted With a Univeral 75% Probability of Success", *Int. Journal Of Modern Physics*, vol. VII, 4 , 2000, pp. 713-720.