文章编号:1001 - 9081(2009)06 - 1520 - 03

重尾分布对网络流量性质的影响

陈 楚,许 勇,张 凌
(华南理工大学 广东省计算机网络重点实验室,广州 510640)
(chuch@scut.edu.cn)

摘 要:重尾边缘分布的网络流量同时存在尺度突发和局部突发。对重尾分布的性质的研究表明重尾分布变量有很强的不稳定性,在网络流量建模中需要足够长的序列来减小重尾分布不稳定性引起的误差。仿真实验结果表明重尾分布的流量在小时间尺度上符合重分形模型。

关键词:重尾分布; 自相似; 重分形; 流量模型
中图分类号: TP393.0 **文献标志码**:A

Influence of heavy-tailed distribution on network traffic

CHEN Chu, XU Yong, ZHANG Ling
(*Guangdong Key Laboratory of Computer Network, South China University of Technology, Guangzhou Guangdong* 510640, *China*)

Abstract: Scale burst and local burst coexist in network traffic. Research shows the heavy-tailed distribution variable is strongly unsteady, which means long enough traffic series is needed in traffic modeling. Simulated results indicate that heavy-tailed network traffic can be appropriately modeled by multi-fractal model in small time scale.

Key words: heavy-tailed distribution; self-similarity; multi-fractal; traffic model

0 引言

对网络流量自相似[1-2]成因的研究,发现网络流大小,连接时间和传输文件大小的重尾分布是导致流量产生自相似的重要原因[3]。重尾分布序列具有无穷大的方差,序列中少量大的观测值对序列性质有重大的影响,网络流量的重尾分布表现为强烈的局部突发。文献[4]用重分形模型描述同时具有尺度突发和局部突发的复杂流量。文献[5]提出用基于alpha-beta 的 ON-OFF 流量模型,其中 alpha 部分可以描述流量的局部突发,beta 部分可以描述流量的长相关。文献[6-7]的研究表明网络流量的边缘分布不是正态的,而是具有重尾性质。本文研究了重尾分布的性质,指出重尾分布的流量具有高可变性和强烈的局部突发,且与重尾分布的尾部指数密切相关。在网络流量建模中应根据流量的重尾程度选择足够长的观测序列来减小高可变性质带来的误差,同时在小时间尺度上应采用重分形模型进行建模。

1 自相似过程与重分形过程

自相似过程有几种不等价的定义,本文采用在网络流量建模中常用的定义[9]。该定义考查平稳随机序列 $X = \{X(i) \mid i \geqslant 1\}$ 的绝对值矩,令:

$$S_m(q) = E \mid X^{(m)} \mid^q = E\left|\frac{1}{m}\sum_{i=1}^{m} X(i)\right|^q; \quad m = 1,2,\cdots \tag{1}$$

若满足以下两个条件,则 X 是自相似的。

1) $S_m(q)$ 与 $m^{\beta(q)}$ 成比例,即:

$$\ln S_m(q) = \beta(q)\ln m + C(q) \tag{2}$$

2) $\beta(q)$ 与 q 呈线性关系。实际上,由于 $X^{(m)}(i) \overset{d}{=} m^{H-1}X(i)$,故:

$$\beta(q) = q(H-1) \tag{3}$$

该方法可以推广到重分形过程。如果 $X(t)$ 的绝对值矩的对数与堆叠度 m 呈对数线性关系,则称为重分形过程。通常用绝对矩法估计序列的 Hurst,如果对于不同的 q 值,Hurst 系数的估计值有明显的差别,则认为序列是重分形的[9]。

对重分形序列的分析是计算其勒让德谱。设 μ 为 R 上有限 Borel 规则测度,对 $x \in \mathbf{R}$,令 $L(x,r)$ 为以 x 为中心,r 为半径的闭区间。定义点 x 的局部测度为:

$$\alpha = \lim_{r\to 0+}\frac{\ln(u(L(x,r)))}{\ln r} \tag{4}$$

α 称为 Holder 指数,它反映了局部奇异程度。对任意的 $q \in \mathbf{R}$ 和 $r > 0$,定义配分函数为测度的 q 次方的求和,即:

$$M_r(q) = \sum u(L(x,r))^q \tag{5}$$

定义质量指数为:

$$\tau(q) = \lim_{r\to 0+}\frac{\ln M_r(q)}{\ln r} \tag{6}$$

重分形的勒让德谱的计算公式为:

$$f_L(\alpha) = \inf_{q\in\mathbf{R}}(q\alpha - \tau(q)) \tag{7}$$

自相似过程的性质之一是慢衰减的方差,其方差的下降速率远远小于它的样本尺寸。这种慢衰减的方差很好地描述了网络流量中的尺度突发,即在所有的时间尺度上,网络流量仍有很强的突发。但自相似过程无法描述网络流量局部的尖峰流量,而重分形过程可同时刻画尺度突发和局部突发,是描述网络流量的理想模型。

收稿日期:2008 - 12 - 12;**修回日期**:2009 - 02 - 20。
基金项目:国家 973 计划项目(2003CB314805); 国家自然科学基金资助项目(60603022)。
作者简介:陈楚(1980 -),男,湖南浏阳人,博士研究生,主要研究方向:网络流量模型; 许勇(1971 -),男,江苏南京人,副教授,主要研究方向:分形几何理论、网络流量模型; 张凌(1962 -),男,江西宜春人,教授,博士生导师,主要研究方向:计算机网络体系结构。

图 1 是短相关序列,自相似序列和重分形序列分别在两个不同的时间尺度下的对比。三个序列的均值都等于 30,均方差都等于 12。其中短相关序列边缘分布是正态的;自相似序列是采用文献[11]中的 FFT 方法生成的 FGN,故其边缘分布也是正态的;重分形序列是采用文献[12]生成的重尾分布的 FARIMA。短相关序列随着堆叠度的增加,突发迅速趋于平缓。自相似序列描述了尺度突发,但从图 1(c)和(d)可以看到,该序列的振幅变化不大,不适合用于描述具有强烈局部突发的流量。重分形的序列则同时描述了尺度突发和局部突发。

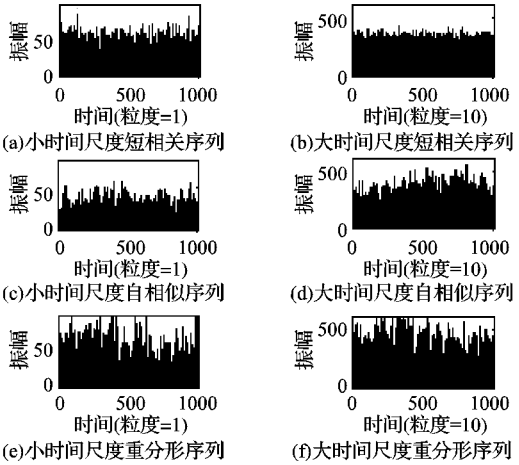


图 1　三种序列的性质对比

2　重尾分布的网络流量

2.1　重尾分布的数学描述

设随机变量 X 的分布函数 $F(x)$ 满足:

$$\overline{F}(x) = 1 - F(x) \sim cx^{-\beta}; 0 < \beta < 2 \tag{8}$$

其中:c 是一个正常数,则 X 是重尾分布的随机变量。X 具有无穷大的方差,尤其当 $\beta \leqslant 1$ 时,X 具有无穷大的均值。

最简单的重尾分布是 Pareto 分布,其概率密度函数与分布函数分别为:

$$p(x) = \beta k^{\beta} x^{-\beta-1}; \ 0 < k \leqslant x \tag{9}$$

$$F(x) = P[X \leqslant x] = 1 - (k/x)^{\beta} \tag{10}$$

其中 k 是随机变量 X 可以取到的最小值。

服从重尾分布的随机变量的特点是:大量的小抽样观察值和少量的大抽样观察值并存,在这些抽样数据集中,虽然大部分抽样值是小的,但是对抽样的均值和方差起决定性作用的是那些少量的大抽样值,这些少量的大抽样取值是不可忽略的。检测数据是否为重尾分布有多种方法,最简单的方法是在对数坐标中画出该数据集的分布函数,如果在一个大的范围内函数是线性递减的,就是重尾分布。对式(8)两边求导可得

$$\lim_{x \to \infty} \frac{\mathrm{d} \ln \overline{F}(x)}{\mathrm{d} \ln x} = -\beta \tag{11}$$

对大的 x 值,重尾分布的分布函数在对数轴上是一条斜率为 $-\beta$ 的直线。

图 2 是正态分布和重尾分布变量的分布函数 log-log 图,两个变量的均值都为 100。重尾分布的尾部指数为 1.8,而按式(11)估算得到的斜率为 -1.82。

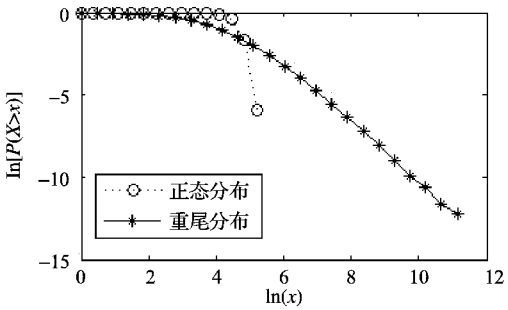


图 2　两种分布的尾部检测

2.2　重尾分布流量的稳定性分析

重尾分布的流量表现出两个特征:一是以很慢的速度收敛到稳态;二是在稳态处表现出高可变性。

重尾分布的慢收敛表现为它的随机变量的抽样平均值收敛的速度非常慢。文献[2]中提出了重尾分布的类中心极限定理,设 X_i 是独立同分布且具有参数 $1 < \beta < 2$ 的重尾分布的随机变量,令:

$$A_n = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{12}$$

$$Z_n = n^{1-(1/\beta)} (A_n - \mu) \tag{13}$$

则当 $n \to \infty$ 时,Z_n 服从 β-稳定分布。β-稳定分布密度函数的形状非常像正态分布,但它有一个幂律参数与原重尾分布相同值的拖尾。

从式(13)可以得出:

$$| A_n - \mu | \sim n^{(1/\beta)-1} \tag{14}$$

可以看出 A_n 以很慢的速度收敛于 μ,当 β 趋近于 1 时,收敛的速度是非常慢的,当 β 等于 1 时,根本就不收敛,此时的均值为无穷大。设用 A_n 去估计均值可达到 k 个数量级的准确率,则:

$$\frac{| A_n - \mu |}{\mu} \leqslant 10^{-k} \tag{15}$$

将式(14)代入(15),可得:

$$n \geqslant c10^{k/[1-(1/\beta)]} \tag{16}$$

即当 n 足够大时,获得 k 位的精确度是可能的。假设 $c = 1$(c 与 μ 有关),表 1 说明了在获得 2 个数量级的精确度所需要的最小抽样数 n,可以看出,当 β 趋近于 1 时,所需要的 n 的数目极大,这在实际中几乎是不可行的。

表 1　重尾分布达到 2 位精度所需的抽样数

β	N
2.0	1E4
1.7	7.2E4
1.4	1E7
1.1	1E22

重尾分布变量的抽样均值即使在稳定处也表现出高可变性。为了更清楚地说明这个性质,设 n 个抽样值的平均为 μ,而下一个抽样值 X 的出现,使得平均抽样值为原来的 $y(y > 1)$ 倍以上。

$$n\mu + X \geqslant (n+1) \times y\mu \tag{17}$$

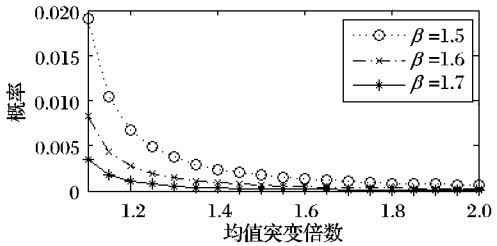
即下一个观测值至少为 $(ny - n + y)\mu$,假设所有的样本均来自均值为 $\mu = k[\beta/(\beta-1)]$ 的 Pareto 分布,当 n 很大时,设上述事件的可能性为 $p_{n\mu}$,则有:

$$p_{n\mu} = P[X > (ny - n + y)\mu] = P[X > n(y-1)\mu] = \left[\frac{k}{n(y-1)k\beta/(\beta-1)}\right]^{\beta} = \left[\frac{\beta-1}{n(y-1)\beta}\right]^{\beta} \quad (18)$$

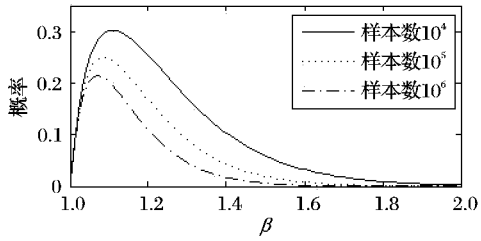
故在 n 次抽样观察中，上述情形至少出现一次的概率为：

$$p = 1 - (1 - p_{n\mu})^n \quad (19)$$

图3(a)是在样本数为 10^5 时样本均值的突变概率。在该样本数下，当β等于1.5时，样本均值突变增加10%的概率约为2%，增加20%的概率约为0.7%。随着β的增大，样本均值突变的可能性明显降低。定义突变的阈值为样本均值增加10%，图3(b)表示了各种情况下突变的概率。从图中可看出，当β大于1.1时，随着β的增大，突变概率显著降低。当β趋于1时，均值接近于无穷大，无法确定抽样均值是否显著增加，突变概率也减小。图3(b)还表明，即使样本数高达10^6，抽样均值仍可能发生显著变化。



(a)样本数为10^5时均值突变分析



(b)样本均值增大10%以上的概率

图3 重尾分布变量抽样均值突变分析

2.3 重尾分布的流量的重分形性质

重尾分布的流量具有强烈的局部突发，用自相似(长相关)模型只能描述其中的长相关而无法描述局部突发。在网络流量建模中，采用绝对矩法确定该流量是否需要用重分形建模。若对不同的q值，绝对矩法估计的Hurst系数值有明显的差别，就表明该流量是重分形的。文献[12]中采用重尾边缘分布的重分形序列描述流量，该模型中的长相关度由参数d控制，自相似参数$H = d + (1/2)$。参数β控制流量的重尾性质，β越小尾部越重，该参数表明的流量局部突发的强烈程度。本文对文献[12]提出的重尾分布的FARIMA序列进行测试，实验结果表明这种方法生成的序列都是重分形的。β越小重分形性质就越明显。

表2 FARIMA序列的Hurst检测结果

尾部参数 β	绝对矩值			
	$q=1$	$q=2$	$q=3$	$q=4$
1.7	0.860	0.790	0.658	0.567
1.8	0.833	0.786	0.669	0.576
1.9	0.807	0.786	0.713	0.626

表2是对3个不同β参数值分别生成100个序列的绝对矩法估计Hurst参数的结果，所有序列的参数d都等于0.3。从表2中可以看出，在不同的q值下，估计得到的Hurst结果差别很大，说明该序列是重分形的。参数β值越小，重分形性质越明显。

图4是利用重尾分布FARIMA序列仿真得到流量的重分形性质分析，其勒让德谱呈右钩形，表明在该流量的各个不同尺度中，尖峰流量的比例很小，小的流量占主要部分，与测量所得的网络流量一致[7]。

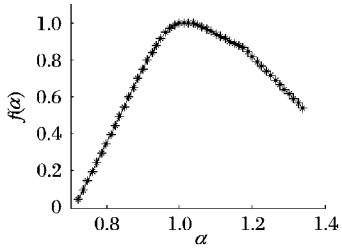


图4 仿真重分形流量重分形谱分析

3 结语

本文分析了重尾分布对流量性质的影响。首先证明了重尾分布具有很强的不稳定性，流量的重尾分布性质要求我们在建模时所测量的序列足够长，尽可能地减少不稳定性带来的误差。其次，本文说明具有重尾边缘分布的序列是重分形的，在流量建模中若测量得到流量的分布具有明显的重尾性质，就应该用重分形进行流量建模。

参考文献：

[1] LELAND W E, TAQQU M S, WILLINGER W, *et al.* On the self-similar nature of Ethernet traffic (extend version) [J]. IEEE/ACM Transactions on Networking, 1994, 2(1): 1-15.

[2] PARK K, WILLINGER W. Self-similar network traffic and performance evaluation [M]. New York: John Wiley & Sons, 2000.

[3] TAQQU M S, WILLINGER W, SHERMAN R. Proof of a fundamental result in self-similar traffic modeling [J]. Computer Communication Review, 1997, 27(2): 5-23.

[4] PATRICE A, RICHARD B, PATRICK F, *et al.* The multiscale nature of network traffic discovery, analysis, and modeling [J]. IEEE Signal Processing Magazine, 2002, 19(3): 28-46.

[5] SHRIRAM S, RUDOLF R, RICHARD B. Network and user driven alpha-beta on-off source model for network traffic [J]. Computer Network, 2005, 48(3): 335-350.

[6] KORN F, MUTHUKRISHNAN S, WU YI-HUA. Modeling skew in data streams [C]// SIGMOD 2006. Chicago: ACM, 2006: 181-192.

[7] DOWNEY A B. Lognormal and pareto distributions in the Internet [J]. Computer Communications, 2005, 28(7): 790-801.

[8] PARK K, WILLINGER W. Self-similar network traffic and performance evaluation [M]. New York: John Wiley & Sons, 2000.

[9] TAQQU MS, TEVEROVSKY V, WILLINGER W. Is the Ethernet data self-similar or multifractal? [J]. Fractals, 1997, 5(1): 63-73.

[10] 肯尼思·法尔科内. 分形几何：数学基础及其应用[M]. 曾文曲，刘世耀，译. 沈阳：东北大学出版社，1991.

[11] PAXSON V. Fast, approximate synthesis of fractional Gaussian noise for generating self-similar network traffic[J]. ACM SIGCOMM Computer Communication Review, 1997, 27(5): 5-18.

[12] STOEV S, TAQQU M S. Simulation methods for linear fractional stable motion and FARIMA using the fast fourier transform [J]. Fractals, 2004, 12(1), 95-121.