自相似网络通信量的滑动平均预测*)

闻 勇 朱光喜
(华中科技大学电子与信息工程系 武汉 430074)

摘 要 研究表明具有重尾特性的自相似性网络通信量表现出长程相关的突发性,对这种不同于传统电话网通信量的长程相关的网络通信量进行描述及预测十分重要。本文在基于对称 alpha-平稳分布过程的自相似通信量模型基础上,提出了两种新的对具有重尾特性自相似网络通信量的滑动平均预测方法。一种是协变正交意义下的线性无偏预测;另一种是双曲线渐近意义下具有对称平稳新息的滑动平均预测,能使预测偏差最小化。对 Bellcore 实验室与 Lawrence 实验室的原始数据进行预测实验,预测结果表明两种预测方法准确有效。

关键词 自相似,alpha-平稳过程,协变正交,滑动平均,预测

Moving Average Forecasting for Network Traffic with Self-Similar

WEN Yong ZHU Guang-Xi
(Department of Electronic and Information Engineering, Huazhong Science Technology University, Wuhan 430074)

Abstract Network traffic with heavy-tailness shows long-range burstiness which is totally different from conventional traffic in telephone network. Characterization and forecast for the long-range dependence traffic is very important for network performance analysis and network design. Two moving average predictors based on alpha-stable self-similar traffic model are presented. First predictor is a linear unbiased estimator based on covariation-orthogonality. Another predictor is asymptotically moving average forecast with symmetrical stable innovation and it can minimize the dispersion. Forecasting experiments for actual traffic trace from Bellcore Laboratory and Lawrence Berkeley Laboratory show that two predictors are accurate and reliable.

Keywords Self-similar, Alpha-stable process, Covariation-orthogonality, Moving average, Prediction

1 引言

自从 Bellcore 实验室[1]发现网络通信量呈自相似(Self-Similar)特征后,各种报告[2]不断证实:广域网和局域网通信量、变比特视频流、ATM 通信流等在大范围的时间尺度上均表现出自相似性。同时研究也表明[3]网络通信量的文件长度、通信量大小、FTP 与 TCP 的传输时间、数据包到达时间间隔、通信量突发时间长短等呈现重尾(Heavy-Tailness)分布,即非高斯(Non-Gaussian)分布。网络通信量的长程相关性是网络通信量的自相似性和重尾特性共同作用的结果,具有重尾分布的自相似网络通信量表现出长程相关(Long-Range Dependent)特性不同于传统电话网通信量中基于泊松过程表现出的短程相关(Short-Range Dependant)特性,即具有长程相关的网络通信量具有更大的突发(Bursty)特性。

近年来,自相似通信量模型不断提出,如:ON-OFF 模型、FARIMA(Fractional Auto-Regression Integrated Moving Average)模型、alpha-平稳模型等,可以在一定程度上对自相似通信量进行模拟。FARIMA 模型可以同时描述网络通信量的长程相关性与短程相关性,但现有的 FARIMA 模型[4]仅针对高斯过程,无法刻画具有重尾特性的自相信网络通信量的突发性。Karasaridis[5]和葛晓虎[6]分别提出了基于 alpha-平稳过程的自相似网络通信量的模型,但他们的模型都有不足之处。网络通信量的预测对于计算机网络的性能分析、结构优化及网络协议设计有重要意义。本文在一种新的基于 alpha-平稳过程的网络通信量模型基础上,给出了两种对于具有重尾特性自相似网络通信量的预测方法:协变正交意义下的滑动平均(moving average)预测;双曲线渐近意义下具有对称平稳新息的滑动平均预测。

文章第 2 节提出新的通信量模型及其参数的估计方法,第 3 节提出两种新的对于具有重尾特性的自相似网络通信量的滑动平均预测方法,第 4 节对实验结果进行分析,最后总结全文。

2 具有重尾特性的自相似性通信量建模

2.1 alpha-平稳分布与 LFSN 过程简介[6]

这一部分给出 alpha-平稳分布和 LFSN(Linear Fractional Stable Noise)过程的基本概念,它们是网络通信量建模的基础。

一维 alpha-平稳分布是典型的重尾分布,它的特征函数 CF(characteristic function)表达式为:

$$E\exp i\theta X=\begin{cases}\exp\{-\sigma^{\alpha}|\theta|^{\alpha}\left(1-i\beta(\sin\theta)\tan\frac{\pi\alpha}{2}\right)+i\mu\theta\} & \text{if } \alpha\neq 1\\ \exp\{-\sigma|\theta|(1+i\beta(\operatorname{sign}\theta)\ln|\theta|)+i\mu\theta\} & \text{if } \alpha=1\end{cases} \quad (1)$$

其中:

$$\operatorname{sign}\theta=\begin{cases}1 & \text{if } \theta>0\\ 0 & \text{if } \theta=0\\ -1 & \text{if } \theta<0\end{cases}$$

*)基金项目:国家自然科学基金重大项目(60496315)。闻 勇 博士生,研究方向:计算机网络。

这里：$\alpha\in(0,2]$为特征指数，表示在分布中突发的程度，当$\alpha=2$时，该分布退化为高斯分布；$\beta\in[1,1]$为偏斜参数，表示整个分布的偏斜程度，与α参数共同决定整个分布函数的形状；σ为尺度参数，表示分布的偏差值；μ为位置参数，表示分布的均值。

随机过程$\{L_{S,H}(t),t>0\}$满足下列条件时为线性分形稳定运动 LFSM(Linear Fractional Stable Motion)：

$$L_{S,H}(t)=\int_{-\infty}^{\infty}(|t-x|^{H-1/\alpha}-|x|^{H-1/\alpha})M_s(dx) \tag{2}$$

这里：a,b是实数常量，$0<\alpha<2$，$0<H<1$，$H\neq\frac{1}{\alpha}$，M是 alpha-平稳随机测度，其存在于具有 Lebesgue 控制测度实数空间中。

线性分形稳定噪声 LFSN 过程 $N_{S,H}(i)$是 LFSM 的平稳增量过程，其定义如下：

$$N_{S,H}(i)=L_{S,H}(i+1)-L_{S,H}(i) \tag{3}$$

LFSN过程是一类 H-sssi(Self-similar with index H and with stationary increments)过程。当 $H>\frac{1}{\alpha}$时，LFSN 过程具有长程相关性。LFSN 过程具有 alpha-平稳过程的基本属性，且具有 H-sssi 过程的基本属性。

2.2 网络通信量的建模

基于 LFSN 过程的网络通信量的模型为：

$$M(i)=c_1N_{\alpha,H}(i)+c_2=c_1(k_d*S^{(\alpha)}_{1,0,0})(i)+c_2 \tag{4}$$

这里：$M(i)$表示在单位时间内到达的包的数量；$N_{\alpha,H}(i)$是参数为$\beta=0,\sigma=1,\mu=0,H>1/\alpha$的 LFSN 过程。模型有四个参数，其物理含义分别为：α参数描示网络通信量的突发程度，H参数描述网络通信量的自相似程度，c_2参数描述网络通信量的平均速率，c_1参数描述网络通信量中相对于平均速率的偏差程度。

根据 alpha-平稳分布的属性[7]知：若$X\sim S^{\alpha}_{\sigma,\beta,0}$，其中$0<\alpha<2$，或$\alpha=1,\beta=0$时，则对于$0<p<\alpha$，必存在常数$c_{\alpha,\beta}(p)$满足：$E(|X|^p)^{1/p}=c_{\alpha,\beta}(p)\sigma$

根据 Hardin 公式[8]：

$$(c_{\alpha,\beta}(p))^p=\frac{2^{p-1}\Gamma(1-\frac{p}{\alpha})}{p\int_0^{\infty}u^{-p-1}\sin^2 u\,du}(1+\beta^2\tan^2\frac{\alpha\pi}{2})^{p/2\pi}\cos(\frac{p}{\alpha}\arctan(\beta\tan\frac{\alpha\pi}{2}))$$

本模型式中：参数$\beta=0$，$N_{\alpha,H}(i)\sim S\alpha S$，$0<\alpha\leqslant 2$，$E|X|^p$为仅由参数$X$中的$\sigma$确定，与$\alpha$、$\beta$无关，故$c_1$可以单独准确描述网络通信量中相对于平均速率的偏差，物理意义明确。Karasaridis 提出的 S4 模型中假定：$\beta=1$，当σ和ρ确定时，$E|X|^p$是一个随α变化的值；葛模型中：$\beta\in(0,1]$，$E|X|^p$随α变化，且是β的偶函数，随$|\beta|$递增。故 S4 模型与葛模型中c_1不能单独准确描述网络通信量中相对于平均速率的偏差程度，故这两个模型理论上有缺陷。

2.3 模型离散化的方法

式(4)的离散表达式：

$$\tilde{N}_{\alpha,H}=(K_d*S^{(\alpha)}_{\sigma,0,0})=\sum_{u=1}^{Mm}K_d(\frac{u}{m})S^{(\alpha)}_{\sigma,0,0}(j-\frac{u}{m}),\quad j=1,\cdots,T \tag{5}$$

这里：$S^{(\alpha)}_{\sigma,0,0}$是 *i. i. d.* alpha-平稳随机变量，$0<\alpha\leqslant 2$，M和m是整数。M定义为记忆，$1/m$网格，$T\leqslant M$为序列长度。核(Kernel)为：

$$K_d(x)=\begin{cases}x^d-(x-1)^d, & x>1\\ x^d, & 0<x\leqslant 1\end{cases}$$

这里：$d=H-\frac{1}{\alpha}$。

2.4 参数估计方法

2.4.1 H 的估计方法：见文[9]。

2.4.2 α和 c_1 的估计方法 由式(1)可得：

$$\log(-\log|\Phi(t)|^2)=\log(2\sigma^{\alpha})+\alpha\log|t|$$

设$y_k=\log(-\log|\Phi(\hat{t}_k)|^2)$，$t_k=\frac{\pi k}{25}$，$k=1,2,\cdots,K$，根据回归由$y_i=m+\alpha w_k+e_k$，求出估值$\hat{\alpha}$和$\hat{m}$后再求解$m=\log(2\hat{\sigma})^{\hat{\alpha}}$，得$\hat{\sigma}$。

当$\alpha\neq 1$时，$S^{\alpha}_{1,0,0}$的 CF 为：

$$\Phi_S(\theta)=E\exp j\theta X=\exp\{-\sigma^{\alpha}|\theta|^{\alpha}\}$$

其 LFSN 逼迫值的 CF，

$$\Phi_{\tilde{N}}(\theta)=\prod_i\Phi_s(h_d(i)\cdot\theta)=\prod_i\exp\{|k_d(i)|^d(-\sigma^{\alpha}|\theta|^{\alpha})\}$$

推导后的模型逼迫值为：

$$\Phi_{\tilde{N}}(\theta)=\exp\{\sum_i|k_d(i)|^d(-\sigma^{\alpha}|\theta|^{\alpha})\}$$

尺度值为：$\sigma=c_1\cdot(\sum_n k^{\alpha}(n))^{\frac{1}{\alpha}}$，则由此可得到：

$$\hat{c}_1=\hat{\sigma}/(\sum_n k^{\hat{\alpha}}(n))^{\frac{1}{\hat{\alpha}}} \tag{6}$$

2.4.3 c_2 的估计方法： 参数c_2可以通过求模型的期望值来得$E[M(i)]=c_1\cdot E[N_{\alpha,H}(i)]+c_2=c_2$，其中$E[M(i)]=u=0$，故参数$c_2$可以用网络通信量的平均值来估计。

3 自相似网络通信量的滑动平均预测方法

3.1 协变正交意义的滑动平均预测

符号幂(signed power $\alpha^{\langle p\rangle}$)的概念：

$$\alpha^{\langle p\rangle}=|\alpha|^p\operatorname{sign}\alpha=\begin{cases}\alpha^p & \text{当}\alpha\geqslant 0\\ -|\alpha|^p & \text{当}\alpha<0\end{cases} \tag{8}$$

对于对称 alpha-平稳过程有协变概念的存在：

假设X_1和X_2是联合$S\alpha S$，$\alpha>1$，且假设Γ是随机矢量(X_1,X_2)的谱测度，X_1在X_2的协变是实数：

$$[X_1,X_2]_{\alpha}=\int_{S_2}s_1s_2^{\langle\alpha-1\rangle}\Gamma(ds) \tag{9}$$

若(X,Y)是联合$S\alpha S$，α和b是实数，则：

$$[aX,bY]_{\alpha}=ab^{\langle\alpha-1\rangle}[X,Y]_{\alpha} \tag{10}$$

设$\{X_i,i=1,\cdots,\infty\}$为 alpha-平稳随机变量，预测误差(error)为：$\hat{e}=\hat{X}_{n+k}-\sum_{i=1}^{n}a_iX_{n+1-i}$，则$[\hat{e},X_{n+1-i}]_{\alpha}=0$为协变正交。若存在$\{a:[\hat{e},X_{n+1-t}]_{\alpha}=0,t=1,\cdots,n\}$，$\hat{X}_{n+k}=\sum_{i=1}^{n}a_iX_{n+1-i}$，使得$E[X_{n+k},aX|X]=0$，则称$\hat{X}_{n+k}$为协变正交线性预测 COLP(covariation-orthogonal linear predictor)。在 COLP 意义下存在无偏估计方法，有下面定理存在：

定理 1 对于所有平稳随机变量X_n，其协变正交一步前向无偏预测为：$E[X_{n+1}|X_n\cdots X_1]=\hat{X}_{n+1}$，它满足

$$\hat{X}_{n+1}=\sum_{i=1}^{n}\theta^{(1)}_{n,i}(X_{n+1-i}-\hat{X}_{n+1-i}) \tag{11}$$

这里：

$$\theta^{(1)}_{n,n-k}=([X_{n+1},X_{k+1}]_{\alpha}-\sum_{i=0}^{k-1}\theta^{(1)}_{n,n-k}[e_i,X_{k+1}]_{\alpha})[e_i,X_{k+1}]^{-1}_{\alpha} \tag{12}$$

(证明见文[10])

COLP 预测的基本原理见文[10]，具体预测算法为：

第一步：对原始数据进行预处理，得到一个均值为零的时间序列；

第二步：对该序列的参数α进行估计；

第三步：根据(12)再求出参数：$\theta^{(1)}_{n,i}$；

第四步：根据式(11)求出数据变化量的预测值，最后求出

原始数据的预测值。

3.2 基于分数差分噪声的滑动平均预测

FARIMA(p,d,q)模型具有下面的形式：

$$\Phi(B)\Delta^d X_n=\Theta(B)\varepsilon_n \tag{13}$$

其中：$\Phi(B)=1+\Phi_1 B+\cdots+\Phi_p B^p$，$\Theta(B)=1+\Theta_1 B+\cdots+\Theta_q B^q$，$B$ 是后向算子，$\Delta^d=(I-B)^{-d}=\sum_{j=0}^{\infty}b_j(-d)B^j$，$d=H-\frac{1}{\alpha}$，$p$ 是自回归阶数，q 是滑动平均阶数。ε_n 为具有无限方差的对称 alpha-平稳新息。

设 $p=0,q=0$ 时，式(13)退化为分数差分噪声(fractional differenced noise)，即：FARIMA(0,d,0)

$$\Delta^d X_n=\varepsilon_n \tag{14}$$

则其解为：

$$X_n=\sum_{j=0}^{\infty}u_j(-d)\varepsilon_{n-j},n=\cdots,-1,0,1,\cdots, \tag{15}$$

其中：

$$\mu_j(-d)=\prod_{k=1}^{j}\frac{k+d+1}{k}=\frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)},j=0,1,2\cdots \tag{16}$$

式(15)也是式(13)的一个解。具有 alpha-平稳新息 FARIMA 的解的存在性与收敛性[7]。

Stuck[11] 提出在对称平稳序列(Symmetric Stable Sequences)的偏差(dispersion)标准：若$\{u_n\}$是独立同分布，其中指数为 α 且如果 $\sum_{j=-\infty}^{\infty}|u_j|^\alpha<\infty$，则 $Y=\sum_{j=-\infty}^{\infty}u_jX$ 也是对称平稳的，事实上：

$$Y\overset{d}{=}\left(\sum_{j=-\infty}^{\infty}|u_j|^\alpha\right)^{\frac{1}{\alpha}}X$$

Y 的偏差为(相对于 X 的)：$disp(Y)=\sum_{j=-\infty}^{\infty}|u_j|^\alpha$

预测系数计算原理见文[11]。具体算法：设 Π 为 $n\times(n+q)$ 阶矩阵，b 为 $1\times(n+q)$ 阶矢量，

$$\Pi=\begin{bmatrix}u_0 & u_1 & u_2 & \cdots & u_q & 0 & & 0\\ 0 & u_0 & u_1 & u_2 & \cdots & u_q & & 0\\ & & & & & \cdots & & \\ 0 & \cdots & 0 & u_0 & \cdots & & \cdots & u_q\end{bmatrix}$$

$$b=(u_0,u_1,\cdots,u_q,0,\cdots,0)$$

又设：$a_0=(\Pi\Pi')^{-1}\Pi b$(a_0X_n 为最小方差预测值)。

再定义：$l_j(a)=[a\Pi_j-b_j]^{a-1}$，$1\leqslant j<n+q$，得到递归公式：

$$a_{k+1}=a_k-(\Pi\Pi')^{-1}\Pi l(a) \tag{17}$$

基于 FDN 的滑动预测算法：

第一步：对原始数据进行预处理，得到一个期望为零的时间序列；

第二步：对该序列的参数 H,α 进行估计，再求出参数：$d=H-\frac{1}{\alpha}$；

第三步：根据式(16)求出系数：$u_i,i=1,\cdots,n$；

第四步：根据式(17)求出预测系数：$a_i,i=1,\cdots,n$；

第五步：根据预测系数求出数据变化量的预测值，最后求出原始数据的预测值。

4 模拟实验及结果分析

网络通信量模型模拟实验中采用 Bellcore 数据 OctExt.TL，模型拟合程度的检验方法主要是采用概率密度函数图和 QQ 图。Alpha-平稳分布无概率密度函数解析表达式，其近似计算方法参见文[12]，QQ 图的原理参见文[13]，实验结果见图 1、图 2。概率密度函数图表明概率密度函数拟合得较好，QQ 图中有少量分位数点有一定的偏差，仅为 1000 个分位数点中的少数，总体上呈直线趋势，这说明本模型对实际数据的拟合程度较好。

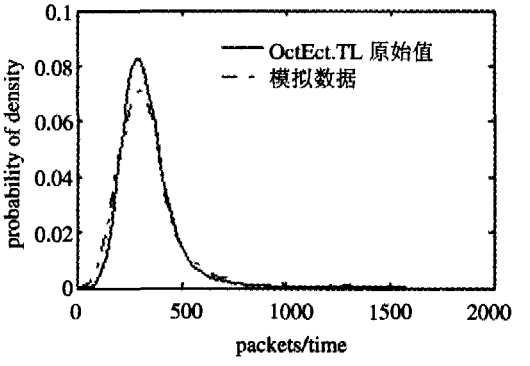

图 1 概率密度图

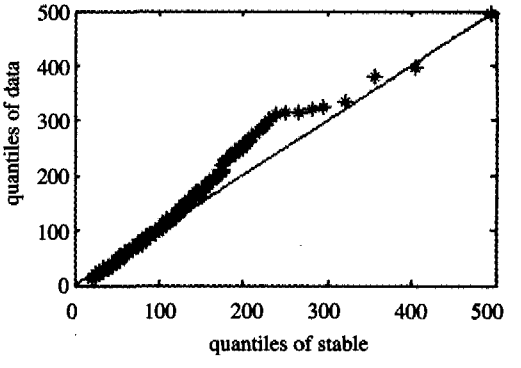

图 2 QQ 图

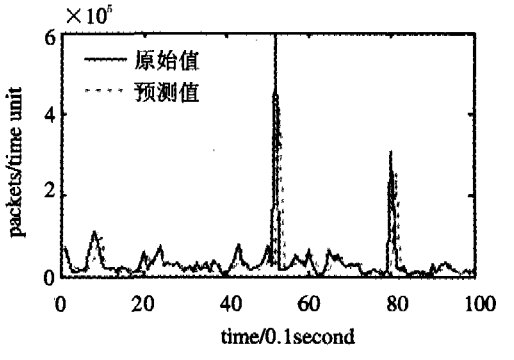

图 3 OctExt. TL 原始数据与预测值

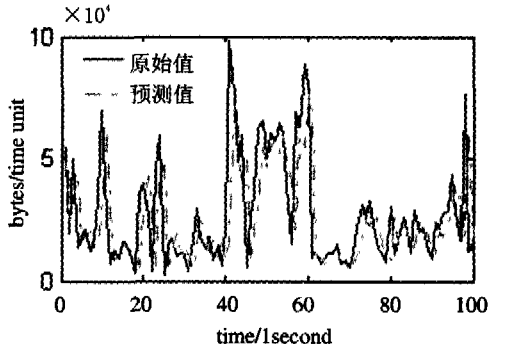

图 4 lbl-tcp-3. tcp 原始数据与预测值

预测实验为一步预测，预测实验数据为 Bellcore 实验室的数据包 OctExt. TL 和 Lawrence Berkeley 实验室的比特流

(下转第 38 页)

in multicast congestion control. In: Doshi B, ed. Proc. of the IEEE INFOCOM. New York: IEEE Communications Society, 1999. 856~863
2 Jiang T, Ammar M, Zegura E. Inter-Receiver Fairness: A Novel Performance Measure for Multicast ABR Sessions. In: Proc. of SIGMETRICS'98, 1998
3 Jiang T, Zegura E, Ammar M. Inter-Receiver Fair Multicast Communication Over the Internet. In: Proc. of NOSSDAV'99, June 1999
4 Jiang T, Ammar M, Zegura E. On the Use of Destination Set Grouping to Improve Inter-receiver Fairness for Multicast ABR Sessions. In: Proc. of INFOCOMM'00, 2000
5 Yang Y, Kim M, Lam S. Optimal partitioning of multicast receivers. In: Proc. IEEE ICNP, 2000
6 Chiu D M, Kadansky M, Provino J, Wesley J, Zhu H. Pruning Algorithms for Multicast Flow Control. In: Proc. of NGC, Palo Alto, 2000
7 Kar K, Tassiulas L. Multicast Rate Control using Lagrangian Relaxation and Dynamic Programming. In: Proc. of the 43rd IEEE Conf. on Decision and Control, Atlantis, Bahamas, Dec. 2004
8 Shenker S J. Fundamental Design Issues for the Future Internet. IEEE JSAC, 1995, 13(7): 1176~1188

(上接第 34 页)

lbl-tcp-3. tcp。图 3、图 4 为采用 COLP 方法的预测结果，预测系数为 200 个；图 5、图 6 为双曲线渐近意义下具有对称平稳新息的滑动平均预测方法的预测结果，预测系数为 50 个。预测结果表明本文提出的两种预测方法均较好地预测原始数据，预测值反映原始值的变化趋势，而且偏差不超过传统标准的 68%。对两个通信量数据集各时间尺度 Trace 进行预测实验，均得到相似的结果。这说明本文所提出的预测算法对于具有重尾特性的自相似网络通信量的广泛时间尺度范围能够准确有效地预测，而且方法简捷快速。尤其是双曲线渐近意义下具有对称平稳新息的滑动平均预测方法避免了传统 FARIMA 模型参数估计的繁琐过程，而且预测结果较好。

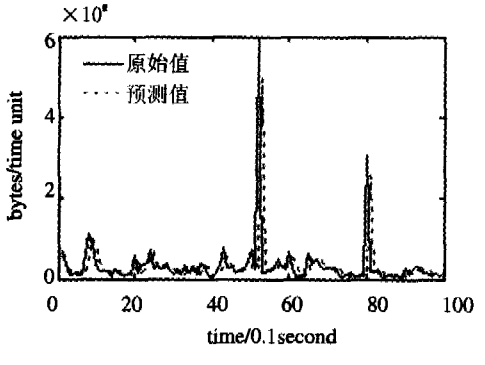

图 5 OctEct. TL 原始数据与预测值

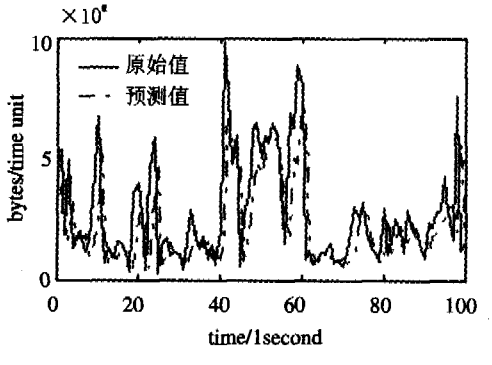

图 6 lbl-tcp-3. tcp 原始数据与预测值

结论 本文在一种新的基于 alpha-平稳过程的网络通信量模型基础上，提出了两种新的滑动平均预测方法：协变正交意义下滑动平均预测方法，双曲线渐近意义具有对称平稳新息的滑动平均预测方法。对不同时间尺度的实际数据的预测结果表明这两种预测方法准确可靠，预测精度随着预测系数个数的增加而有提高。这两种对于具有重尾特性的自相似网络通信量的预测方法是本文的创新之处。

目前，我们正努力将新的网络通信量预测方法应用于网络性能分析与网络管理中。

参 考 文 献

1 Leland W, Taqqu M, Willinger W, et al. On the self-similar nature of Ethernet traffic. In: Proc. ACM SIGCOMM '93, 1993. 183~193
2 Paxson V, Floyd S. Wide-area traffic: The failure of Poisson modeling. In: Proc. ACM SIGCOMM '94, 1994. 257~268
3 Crovella M E, Taqqu M S, Bestavros A. Heavy-tailed probability distributions in the World Wide Web. In: A practical guide to heavy tails, chapman&Hall, 1998
4 Shu Y, Jin Z, Wang J, Yang O W. Prediction-based admission control using FARIMA models. In: Proc. IEEE ICC'00, New Orleans, USA, June 2000, 3: 1325~1329
5 Karasaridis A, Hatzinakos D. Network heavy traffic modeling using α-stable self-similar processes. IEEE Transactions on Communications, 2001, 49(7)
6 Ge X H, Zhu G Xi, Zhu Y T. An improved modeling of network traffic using alpha-stable processes. The Chinese Journal of Electronics, 2003, October
7 Samorodnitsky G, Taqqu M. Stable non-gaussian random processes. Chapman and Hall, NewYork, London, 1994
8 Hardin J C D. Skewed stable variables and processes; [Technical Report 79]. Center for Stochastic Processes at the University of North Carolina. Chapel Hill, 1984
9 Clegg R G. A Practical Guide to Measuring the Hurst Parameter N. In: Proc. of 21st UK Performance Engineering Workshop, School of Computing Science Technical Report Series, CS-TR-916, University of Newcastle, 2005. ISSN 1368-1060
10 Hill J. Minimum dispersion and un-biasedness; 'best' linear predictors for stationary ARMA α-stable processes. Working paper No. 00. 06, September 2000. center for economic analysis department of economics. University of Colorado at Boulder
11 Stuch B W. Minimum error dispersion linear filtering of scalar symmetric stable processes. IEEE Trans. Aut. Cont. 1978, 23: 507~509
12 Cline D B H, Brockwell P J. Linear prediction of ARMA processes with infinite variance. Stochastic Processes and their Applications North-Holland, 1985, 19: 281~296
13 Zolotarev V M. On representation of densities of stable laws by special function. Theory of Probability and Its Applications, 1994, 39: 354~362
14 Rice J A. Mathematical statistics and data analysis, Second Edition. Wadsworth Inc, 1995