自相似网络流量预测的分析和研究

王西锋[1,2],高 岭[1],张晓孪[2]
(1.西北大学 信息科学与技术学院,陕西 西安 710127;
2.宝鸡文理学院 计算机科学系,陕西 宝鸡 721007)

摘 要:网络流量模型能准确和全面地刻画流量数据的各种统计特性,因而成为网络研究的热点。讨论了传统模型预测的弊端,描述了网络流量自相似的基本特征,分析了三个重要自相似模型的特点和存在的问题,使用实际网络流量验证了流量的自相似性并结合多分形小波模型对网络流量进行预测,探讨了流量模型预测的新技术和进一步研究的问题。

关键词:自相似;长相关;重尾分布;小波分析;流量模型
中图分类号:TP393.06 **文献标识码**:A **文章编号**:1673-629X(2007)11-0042-04

Analysis and Research on Self-Similar Network Traffic Forecast

WANG Xi-feng[1,2],GAO Ling[1],ZHANG Xiao-luan[2]
(1.Information Science and Technology Institute,Northwest University,Xi'an 710127,China;
2.Department of Computer Science,Baoji College of Arts and Science,Baoji 721007,China)

Abstract:Network traffic models can concisely and accurately describe kinds of traffic data statistical characteristics,so it became a hot topic of network capability research. This paper discusses the drawback of the traditional traffic models,describes the basal character of self-similar and analyzes detailedly the feature and existing problems of three important self-similar traffic models. Afterwards,validate the self-similar and forecast the network traffic with the real network traffic data. The latest technologies and the next research problems are given in the end.

Key words:self-similar;LRD;heavy-tail distribution;wavelet analysis;traffic model

0 引 言

随着网络带宽的迅速增加和各种网络服务的广泛应用,网络规模不断扩大,体系结构愈发复杂,对其运行控制及管理维护日趋困难。网络流量模型是进行网络性能评价和网络规划的基础,能够反映网络的真实流量状况,人们通过模型预测也可以认识和分析网络流量数据的变化规律及行为特征。因此,在通信网络技术发展的过程中,针对网络流量进行建模和预测的研究一直备受人们关注。

20世纪70年代人们主要借鉴公共交换电话网络(PSTN)的流量模型,使用Poisson模型来描述数据网络,随后又引入了Markov模型和其它的随机模型,这些传统模型的共同特点就是所描述的流量时间序列都具有短相关性。随着网络节点数的指数增加和新应用的不断出现,流量特征发生了显著变化。90年代初期,文献[1]首先证实了局域网流量中存在自相似性,文献[2]证明了WWW的广域网流量的自相似和突发性,文献[3]也发现了WWW流量具有自相似性。通过对大量实际网络流量的分析,研究人员发现不论是局域网还是广域网,网络流量数据都呈现出明显的自相似特性。

收稿日期:2007-01-04
基金项目:陕西省自然科学基金项目(2005F36)
作者简介:王西锋(1978-),男,陕西渭南人,硕士研究生,研究方向为计算机网络性能分析;高 岭,教授,研究方向为计算机网络性能分析。

1 网络流量的自相似特性

1.1 自相似的数学描述

人们把实际网络中存在的流量突发性不随时间轴变化而变化的特性称为自相似,它意味着流量数据的局部以某种方式与整体相似。贝尔实验室对自相似过程的数学描述如下[1]:

考察一个广义平稳过程 $X=\{X_t:t=0,1,2,\cdots\}$,其中 X_t 表示网络中第 t 个单位时间内到达的网络业务数目,平稳时间序列的期望和方差分别为:$\mu=E[(X_t)]$,$\delta^2=E[(X_t-\mu)]$,自相关函数为:$r(k)=E[(X_t-\mu)(X_{t+k}-\mu)]/\delta^2$。对每个 $m=1,2,3\cdots$,令

$X_k^{(m)} = (X_{km-m+1} + X_{km-m+2} + \cdots + X_{km})/m, k = 1, 2, 3\cdots, x^{(m)} = \{X_k^{(m)}, k = 1,2,3\cdots\}$ 是一个根据 X 得到的 m 阶聚合序列,记 $r^{(m)}$ 为 $X^{(m)}$ 的自相关函数。

如果对所有的 $m = 1,2,3\cdots$,都有 $r^{(m)}(k) = r(k) \to k^{-\beta}, m \to \infty, k = 0,1,2\cdots$,称 X 为渐进二阶自相似过程,其中 $H = 1 - \beta/2$ 为自相似参数,取值于 0.5 与 1.0 之间,用来刻画自相似程度的强弱。

如果 $r(k)$ 满足 $\sum r(k) = \infty$,称 X 具有长相关性,物理意义指当前流量值与其所有的历史值都有关。在渐进二阶自相似的情况下,自相似就意味着流量数据具有长相关结构。

1.2　自相似的成因分析

对网络流量数据的分析和研究发现自相似的产生不是单一因素的结果,而是多种因素共同作用的结果。根据实际观测和理论分析[2],人们发现在微观上具有重尾特性的分布能够在宏观上产生明显的长相关、自相似性,即自相似的产生和重尾分布有紧密联系。重尾分布的定义如下:

随机变量 X 的分布是重尾分布当且仅当:$1 - F(x) = P[X \geqslant x] \to cx^{-\beta}, \beta \geqslant 0, x \to \infty$。

有学者认为由于网络中某些参数(如业务源和应用层的行为、文件大小、传输时间等)服从重尾分布,从而导致网络流量在时间尺度上的突发性和自相似[3];也有学者认为是因为传输层协议的相互影响导致了自相似,因为 TCP 协议占整个网络流量的 90% 左右,而 TCP 协议的超时重传机制和指数后退机制使得分组到达间隔时间呈重尾分布,从而使得流量呈现出自相似性[4];Willinger 等则认为是大量独立的 ON/OFF 数据源在流量汇聚的过程中导致了网络流量的自相似性[5]。

2　自相似网络流量模型

对于具有自相似特性的网络,由于传统模型仿真的流量通常在时域仅具有短相关性,经过时间上的平均后突发性会趋于平稳状态,不能对网络流量进行准确描述和预测;而基于自相似特性的自相似模型却可以较好 F 描述实际网络流量。下面给出几种重要的自相似模型。

2.1　ON/OFF 模型

ON/OFF 源叠加模型是对传统模型的扩展,将自相似过程看成是无数用户数据源叠加的结果。模型定义了大量的数据源,每个源有 ON 和 OFF 两个状态,各个数据源相互独立且状态持续时长符合重尾分布。当数据源处于 ON 状态时以恒定的速率产生数据,而处于 OFF 状态时则不产生任何数据。自相似过程可以通过叠加大量的具有重尾分布的 ON/OFF 数据源得到,当数据源的个数趋于无穷时,总的网络流量是趋于渐近自相似的。

ON/OFF 模型有明确的物理意义,把复杂的聚合流量特征分析细化到对每个信号源的分析,可以解释产生自相似的部分原因;缺点是这种自相似业务只是渐近自相似的,且假设前提过于严格,大多数网络业务的分布是无法建立在此前提上的。

2.2　分形布朗运动模型(FBM)

FBM 是 Manderbrot 和 Van ness 在 1968 年提出的一种统计自相似过程的数学模型,Norros 在文献[6]中讨论了使用分形布朗运动对网络流量进行建模预测。FBM 是一种结构简单的建模方法,令 A_t 表示在时间 $(0,t)$ 中分组到达的个数,Z_t 为标准 FBM 生成的随机变量,只需要流量平均速率 m、网络流量方差 a 和 Hurst 参数 3 个参数就可以通过 $A_t = mt + \sqrt{am}Z_t$ 表达式完整地刻画整个模型。产生分形布朗运动的快速算法是随机中点置位法,它是产生自相似过程速度最快的算法,计算的复杂度仅为 $O(n)$。

FBM 是使用最广泛的自相似模型之一,在数学上有坚实的理论基础,常用于进行理论分析;但 FBM 是严格的自相似过程,不能对具有短相关性的流量很好地进行分析。因此,分形布朗运动模型并不能完整地描述网络实际情况。

2.3　多分形小波模型(MWM)

自相似网络流量是非平稳的时间序列,而小波分析是一种变分辨率的时频联合分析方法,通过伸缩和平移运算由粗到细地观察信号;同时小波变换是一种不损失任何信号信息的守恒变换,能很好地描述和处理非平稳时间序列。因此,近年来小波技术在网络流量建模和预测中的应用越来越广泛。

多重分形小波模型[7]是基于 Haar 小波的网络流量模型,这里 Haar 小波函数和尺度函数构成了一个简单的小波正交基,其尺度函数和母函数如下:

$$\phi(t) = \begin{cases} 1, 0 \leqslant t < 1 \\ 0, \text{else} \end{cases} \tag{1}$$

$$\Psi(t) = \begin{cases} 1, & 0 \leqslant t < 0.5 \\ -1, & 0.5 \leqslant t < 1 \\ 0, & \text{else} \end{cases} \tag{2}$$

Haar 小波变换的尺度系数和小波系数可通过下面的递推公式计算:

$$\begin{cases} a_{j-1,k} = 2^{-1/2}(a_{j,2k} + a_{j,2k+1}) \\ d_{j-1,k} = 2^{-1/2}(d_{j,2k} + d_{j,2k+1}) \end{cases} \tag{3}$$

即,

$$\begin{cases} a_{j,2k} = 2^{-1/2}(a_{j-1,k} + d_{j-1,k}) \\ a_{j,2k+1} = 2^{-1/2}(a_{j-1,k} - d_{j-1,k}) \end{cases} \quad (4)$$

由于网络流量数据都是非负的,要模拟网络流量,就必须保证尺度系数和小波系数是非负的。尺度系数表示流量在不同尺度变换下的逼近值,因此流量值非负时,尺度系数值一定非负。即,

$$\forall j,k, f(t) \geqslant 0 \Leftrightarrow a_{j,k} \geqslant 0 \quad (5)$$

为保证流量值非负,MWM 模型中引入因子 $A_{j,k}$,并使得

$$a_{j,k} = A_{j,k} * d_{j,k} \quad (6)$$

这里 $A_{j,k}$ 是[-1,1]之间的独立随机变量,由式(4)可知,只要将 $A_{j,k}$ 限定在[-1,1]的区间内就能保证模拟流量值非负。MWM 模型产生模拟数据序列的过程可以简要描述如下:

(1)设 $j = 0$,计算大尺度的系数 $a_{0,0}$,建立起流量的全局均值;

(2)在尺度 j 上,产生随机倍乘变量 $A_{j,k}$,选 $A_{j,k}$ 为对称分布的 β 分布,并通过式(6)计算,$a_{j,k}, k = 0, \cdots, 2^j - 1$;

(3)在尺度 j 上,由式(4)计算出尺度 $j+1$ 的 $d_{j+1,2k}$ 和 $d_{j+1,2k+1}, k = 0, \cdots, 2^j - 1$;

(4)增加 j,重复步骤(2)、(3),直至达到尺度 $j = n$ 为止。

MWM 是一个乘法模型,能够对网络流量的短相关和长相关特性进行描述,因而可以很好地匹配实际网络流量;不足的是小波变换系数并非在每个尺度下都独立,而且小波基的选取也影响模型的质量,基于小波预测的效果现在并不能让人满意;但由于小波分析是处理非平稳序列最有效的方法,流量预测有望应用小波技术取得突破。

3 实际流量的自相似性分析和预测

3.1 实际流量的自相似性

为了确信流量的自相似性,对实际的网络流量数据进行了验证。实验数据来源于流量文库:http://newsfeed.ntcu.net/news/2006/,收集了主节点路由器 NEWS 从 2006 年 11 月 1 日到 12 月 10 日,共 40 天的网络每小时通过流量,总计得到 $24 \times 40 = 960$ 个实验数据,从而形成原始网络流量时间序列 $\{f(t), t = 1,2,\cdots,960\}$,如图 1 所示。

图 1 中最大流量数据为 5324.8MB/h,最小仅为 504.6MB/h,两者相差很大,说明了实际网络流量是不平稳的时间序列,具有很强的突发性;网络流量在时间尺度上呈现出明显的自相似性,局部和整体流量数据波形类似;同时也观察到,网络流量在每天、每周都大致呈现出一种周期性的变化规律,这也是在进行流量建模时应该考虑的。

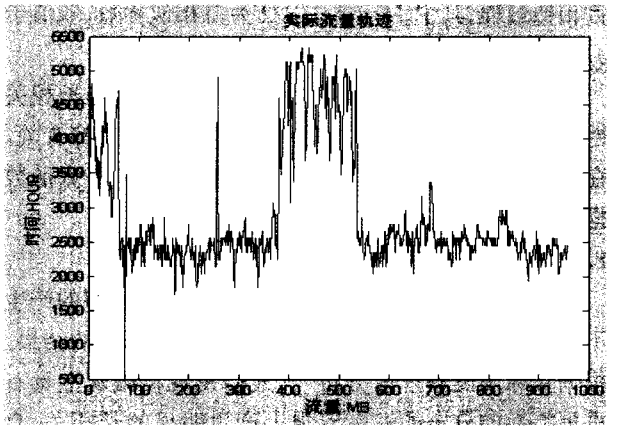

图 1 原始流量序列

自相似参数 H 的估计方法很多,常用的大都是图形化或半图形化的方法,如方差/时间图、R/S 图、Whittle 估计方法和小波估计方法,得到流量序列的 H 参数后,就可以用来判断网络流量是否具有自相似性。笔者应用方差/时间图参数估计法,取 20 个不同的时间单位,分别计算其方差 V_m,然后对 $\ln(m)$ 和 $\ln(V_m)$ 做线性拟合,通过估计得到该时间序列的自相似参数 $H = 0.765$,表明该流量具有非常明显的自相似性。

3.2 实际流量的小波预测

为了对小波分析在流量预测中的效果进行分析,使用前面的多分形小波模型对网络流量进行预测。由于连续小波变换是冗余变换,这里使用离散小波变换来处理。应用 MWM 模型对网络流量进行预测,仿真流量如图 2 所示。

图 2 是仿真的预测流量和实际流量,其中横坐标

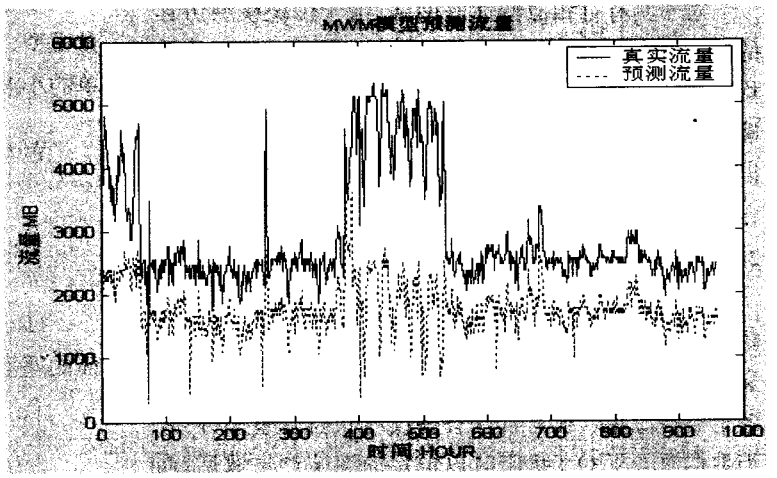

图 2 MWM 预测流量

为时间段,纵坐标为网络流量。与实际流量数据进行比较,可以看出整个预测流量在变化趋势、变化快慢和分散程度上都较好地反映了网络实际流量。因此,小波分析对自相似网络流量的预测性能是比较好的。近年来也不断有学者结合小波技术对网络流量进行建模预测,都取得了不错的预测效果[8,9]。

4 结束语

网络流量的自相似性决定了网络的行为特征,只有基于网络重要特征—自相似的建模才能准确描述网络实际情况。文中对主要的自相似网络模型预测方法进行了分析和总结,并应用多分形小波模型对网络流量进行了验证和预测,表明实际网络流量确实具有自相似性。

网络流量行为随着时间和地域的不同会呈现出很大的变化,因此想提供一种针对网络流量的通用化预测模型是很困难的。今后的流量预测会考虑实际网络流量存在的长、短相关性,根据不同网络分量的特点分别选用合适的方法进行预测,然后合成得到网络预测流量。

近几年,有研究人员也将自相似流量和混沌理论、模糊理论和神经网络理论结合起来研究网络的行为特性,这些新理论的引入都将对流量预测产生重要的影响。

参考文献:

[1] Leland W E, Taqqu M S, Willinger W, et al. On the Self - similar Nature of Ethernet Traffic (Extended Version)[J]. IEEE/ACM Transactions on Networking, 1994, 2(1): 1 - 15.

[2] Paxson V, Floyd S. Wide - area traffic: the failure of Poisson modeling[J]. IEEE/ACM Transaction on Networking, 1995, 3(3): 226 - 244.

[3] Crovella M, Bestavros A. Self - similarity in World Wide Web traffic: evidence and possible causes[J]. IEEE/ACM Transactions on Networking, 1997, 5(6): 160 - 169.

[4] 赵佳宁,李忠诚. 基于模拟的网络流量自相似现象分析[J]. 计算机科学,2001,28(11):57 - 61.

[5] Willinger W, Taqqu M S, Sherman R, et al. Self - similarity through high - variability: statistical analysis Ethernet LAN traffic at the source level[J]. IEEE/ACM Transactions on Networking, 1997, 5(1): 71 - 86.

[6] Norros I. On the Use of Fractional Brownian Motion in the Theory of Connectionless Traffic[J]. IEEE JSAC, 1995, 13(6): 953 - 962.

[7] Riedi R H, Crouse M S, Ribeiro V J. A multiscale wavelet model with application to network traffic[J]. IEEE Trans. on Info. Theory, 1999, 45(3): 992 - 1018.

[8] 李　捷,刘瑞新,刘先省,等. 一种基于混合模型的实时网络流量预测算法[J]. 计算机研究与发展,2006,43(5):806 - 812.

[9] 洪　飞,吴志美. 基于小波的多尺度网络流量预测模型[J]. 计算机学报,2006,29(1):166 - 170.

(上接第 38 页)

部分仍有待提高。另外对搜索的结果可以进行优化、分类,双管齐下,能最大限度地提高引擎效率,体现个性化服务的特点。

参考文献:

[1] 陈笑辉,范晓红. 搜索引擎的分类体系及性能评价[J]. 中国信息导报,1999(7):18 - 20.

[2] 朱俊卿. 搜索引擎 google 研究[J]. 现代图书情报技术,2001(4):34 - 36.

[3] 武助宇. 中文搜索引擎的发展现状、问题与对策[D]. 湘潭:湘潭大学,2002.

[4] 王开铸. 自然语言理解[M]. 哈尔滨:哈尔滨工业大学出版社,1996.

[5] 姚天顺. 自然语言理解,一种让机器懂得人类语言的研究[M]. 北京:清华大学出版社,2002.

[6] 袁占亭,张爱民,张秋余. 基于概念的 Web 信息检索[J]. 计算机工程与应用,2003(36):173 - 175.

(上接第 41 页)

的识别都有着很好的借鉴作用。

参考文献:

[1] Lee Y H, Kassam S A. Generalized median filtering and related nonlinear filtering techniques[J]. IEEE Transactions on Acoustics, Speech, Signal Processing, 1985, 33(3): 672 - 683.

[2] 胡小峰,赵　辉. Visual C + + /MATLAB 图像处理与识别实用案例精选[M]. 北京:人民邮电出版社,2004.

[3] 求是科技. Visual C + + 数字图像处理典型算法及实现[M]. 北京:人民邮电出版社,2006.

[4] Gonzalez R C, Woods R M. Digital Image Processing[M]. 2nd ed. 北京:电子工业出版社,2002.

[5] Liu Ji - lin, Song Hong - tao, Ding Li - ya. Vehicle License Plate Recognition System with High Performance[J]. 自动化学报,2003,29(3):457 - 465.

[6] 夏　勇,田　捷,邓　翔. 一种高效的自适应指纹图像压缩算法[J]. 计算机学报,1999,22(5):525 - 528.

[7] 王树禾. 图论及其算法[M]. 合肥:中国科学技术大学出版社,1990.