大规模网络流量行为累加分解研究

程 光　龚 俭

（gcheng,jgong@njnet.edu.cn　东南大学计算机系 南京 210096）

摘要：在大规模网络中的流量行为体现为一个相当复杂的非线性，目前国内外对它的研究还没有成熟的方法。本文采用一种累加模型将复杂大规模网络流量分解成趋势项、周期项、随机项。根据这一分解，利用不同的数学工具分别建模三个相对简单的子系统来仿真复杂流量。为了检查我们的模型，使用分解模型分析 CERNET 主干网络长期流量行为，并将分析结果同传统的 ARIMA 季节模型比较，结果表明累加模型在描述流量宏观行为时具有简单和高精度的优点。

关键词：累加分解模型、流量行为、非线性、预测

分类号： TP393

1 引言

大规模网络管理、规划、设计以及新一代网络体系结构的设计等均离不开对网络流量行为（traffic behavior）的理解。研究网络流量行为首先是要直接对测量流量资料进行统计分析，寻找统计规律，这方面代表性的工作是 MCI 的 Thompson[1]，他完成了对 MCI 网络两个测量点的 24 小时、7 天的流量进行详细分析。其次是在流量统计分析的基础上建立流量模型，如：94 年 Leland[2]等人对以太网测量资料进行统计分析发现以太网流量具有自相似性，并建立网络流量自相似模型。对网络流量行为特征的研究还可在不同测量时间粒度上展开。Paxson 和 Floyd[3]的研究发现，不同时间粒度流量服从不同的行为规律：毫秒级超细时间粒度的流量行为由于主要受网络协议的影响，因而不体现自相似特征；小时级以上粗时间粒度的流量行为由于主要受外界因素的影响，也不具有自相似性，而是一种非线性复杂的过程；而秒级细时间粒度的流量行为体现出自相似性。本文的研究是粗时间粒度下流量时间序列模型，其结果更多地体现网络行为的宏观特征，因此也称为宏观流量行为。

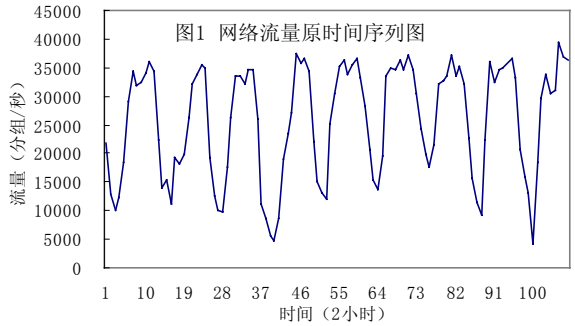
在描述网络流量行为的模型中，时间序列模型起着相当重要的作用。传统的宏观流量时序模型只能处理平稳过程和特殊的非平稳过程，如：AR[4]模型、ARMA[5]模型用于解决平稳过程，ARIMA[5]模型和 ARIMA 季节模型[6]用于处理齐次的非平稳性过程等。由于大规模网络本身是复杂非线性系统，同时又受多种复杂外界因素的影响，其宏观流量行为往往复杂多变，资料中既含有多种周期类波动，又呈现非线性升、降趋势，还受到未知随机因素的干扰，而这些特点不能用传统模型来描述。本文根据网络流量特点，使用累加运算将复杂流量系统分离成结构相对简单的子系统，通过对各子系统的分别研究（建模、预测等）来获得流量行为的整体信息，并描述和预测流量行为的非线性规律。

2 流量行为的基本分析

2.1 非线性流量时序的结构特点

图 1 是 2001 年年初 CERNET 华东(北)地区网络中心对 CERNET 华东(北)地区网与

CERNET 主干网交换流量的监测资料，采样的时间尺度是 2 小时，图中表明，流量非线性序列中含有以下成分：



（1）趋势成分 T_t：该成分显示了流量序列的长期行为规律，主要呈现增长、下降和平稳三类特性。前两种特性可用多项式、指数函数等描述，平稳性的结构特性需要时序建模。

（2）周期成分 C_t：流量序列往往含有不只一个周期，其振幅亦可能随时间而增大或衰减。可通过滑动和滤波，也可用三角函数的组合来描述。

（3）随机成分 E_t：由随机因素影响而导致的波动无确定规律，统计特性多为零均值的白噪声或平稳序列。

将流量时间序列分解称三个组成部分之后，就可按各组成项的行为分别进行研究以获得对流量总体行为的认识。

2.2 流量系统分解

累加分解模型按并联法模式进行流量分解。并联法模式的系统由几个结构较简单的子系统并联而成。图 1 的{ Z_t }可用并联法来模拟：

$$Z_t = T_t + C_t + x_t$$

式中 T_t 为趋势项、C_t 为周期项、x_t 为平稳序列，可用 ARMA 模型拟合。

将系统输出分解为如下子序列的模型：

$$Z_t = T_t + C_t + E_t \qquad (1)$$

统称为加法模型，网络流量行为可采用此类模型拟合和预测。加法模型有多种分解的方法和途径，本文主要讨论用累加季节调整滤波来分解流量非线性结构。

3. 累加分解模型

本文中提出的累加法模型中使用累加平均法来分离流量资料中趋势（T）、周期（C）及随机因子（E）。下面我们先研究累加平均法的性质和定义。

3.1 累加平均定义

设序列为 $\{x_t\}$, t=1, 2, …, N, 定义

$$MA_S(x_t) = \frac{1}{S}\sum_{i=0}^{S-1} x_{t+i}\ ,\ t=1,\ 2,\ \ldots,\ N\text{-}S\text{+}1 \qquad (2)$$

为序列$\{x_t\}$的跨度为 S 的对称加法平均，简记为 MA_S。

根据（2）式对序列$\{x_t\}$进行运算，将输出序列记为$\{y_t^*\}$，$t^*=(S-1)/2+t$，是 S 个输入值 x_t, …, x_{t-S+1} 的中心时刻，由此称（2）式为对称加法平均或中心加法平均。S 为奇数时，t^*为整数；S 为偶数时，t^*非整数。将上两种加法平均分别记为 MA_{2m+1} 和 MA_{2m}。

MA_S 能成为分离流量资料的有效工具是由其性质决定，MA_{2m+1} 具有如下重要性质：（1）A_{2m+1} 对直线序列“透明”。即若输入$\{x_t\}$满足 $x_t=a+bt$, t=1, 2, …, N，则输出 y_t^*满足 $y_t^* = MA_{2m+1}(x_t) = MA_{2m+1}(a+bt) = a+b(t+m) = x_{t+m}$，即{ y_t^*}与$\{x_t\}$在对应时刻上的值相等。

（2）若输入序列{x_t}为周期序列，满足 $\frac{1}{2m+1}\sum_{i=0}^{2m} x_{t+i} = \mu$，t=1,2,…,N-2m。μ 为常数，则输出的 y_t^*亦为常数 μ 。

（3）若{x_t}独立同分布序列，方差为 $Varx_t = \sigma^2$，则 $Vary_t = \frac{\sigma^2}{2m+1}$。

从以上三个性质表明，MA_{2m+1} 能压缩随机波动，分离周期因素，保持资料趋势。由于 MA_{2m} 的输出序列相应时刻非整数，而 MA_{2m+1} 又只能分离奇数周期，这使得实际应用不便，为克服这一困难，本文中采用 2 阶累加平均法。设 y_{t1}, $t_1=(S_1-1)/2+t$, t=1, 2, …, $N-S_1+1$ 是序列{x_t}, t=1, 2, …, N 的 MA_{s1} 的输出，定义 2 阶对称加法平均，记为 $MA_{S2\times S1}$。

$$Z_{t_2} = MA_{S_2}(y_{t_1}) = MA_{S_2}(MA_{S_1}(x_t)) = \frac{1}{S_2 S_1}\sum_{i=0}^{S_2-1}\sum_{j=0}^{S_1-1} x_{t+i+j} \qquad (3)$$

其中 t=1, 2, …, $N-S_1-S_2+2$; $t_1=(S_1-1)/2+t$; $t_2=(S_2-1)/2+t_1$。

3.2 估计趋势模型

在该模型中我们采用 $MA_{2\times S}$ 累加平均模型进行流量序列中的趋势项分解，$MA_{2\times S}$ 中 S 是时间序列中的周期长度，下面考察 $MA_{2\times S}$ 模型具有的性质。

$$Z_{S/2+t} = MA_{2\times S}(x_t) = \frac{1}{2\times S}\left(\sum_{i=0}^{S-1} x_{t+i} + \sum_{i=0}^{S-1} x_{t+i+1}\right)，\ t=1,2,\ldots,N-S \qquad (4)$$

$MA_{2\times S}$ 用来分离流量序列中的季节因子以获取趋势 T_t 的估计，且 C_t 以 S 为周期并满足 $\sum_{i=0}^{S-1} C_{t+i} = 0$，则 $MA_{2\times S}(x_t) = MA_{2\times S}(T_t) + MA_{2\times S}(C_t) + MA_{2\times S}(E_t)$

$$= MA_{2\times 12}(T_t) + MA_{2\times 12}(E_t)$$

由 3.1 节性质（3）知 $MA_{2\times S}(E_t)$ 值很小。如果 T_t 局部线性，则由性质（1）知 $MA_{2\times S}(x_t)$ 即可作为的 T_t 估计。使用 $MA_{2\times S}$ 模型要求 S 为偶数；如果 S 为奇数，则直接用 MA_{2m+1} 来估计趋势项，其中 2m+1=S。在本文中，因我们的采样尺度是 2 小时，流量序列是以天为周期，所以 S＝12，故使用的模型是 $MA_{2\times 12}$。

3.3 估计周期模型

为了能估计流量序列中的周期成分，先讨论 $MA_{3\times 3}$ 的输入输出关系：

$$Z_{2+t} = MA_{3\times 3}(x_t) = \frac{1}{3}(y_{1+t} + y_{2+t} + y_{3+t}) = \frac{1}{9}(x_t + 2x_{t+1} + 3x_{t+2} + 2x_{t+3} + x_{t+4}) \qquad (5)$$

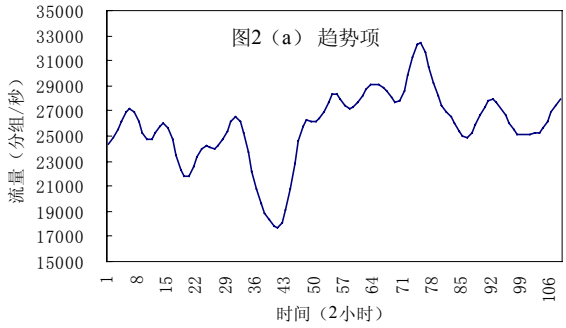
$MA_{3\times 3}$ 用来估计季节因子。设欲从周期为 S 的流量时间序列 $x_t=C_t+E_t(\{x_t\})$中（不含趋势项 T 或已被分离）中估计周期季节因子。如第一个周期的资料 $x_{S\times r+1}=C_{S\times r+1}+E_{S\times r+1}$，做如下滑动平均 $MA_{3\times 3}(x_{S\times r+1}) = MA_{3\times 3}(C_{S\times r+1}) + MA_{3\times 3}(E_{S\times r+1})$，理想情况下，$MA_{3\times 3}(E_{S\times r+1})$ 较小，$C_{S\times r+1}$ 因不同周期同一时刻的资料变动不大，故可用来作为 $C_{S\times r+1}$ 的估计。用 $MA_{3\times 3}$ 而不用 MA_3 是因为前者比后者更能减少随机因子的影响，对周期因子估计更精确。

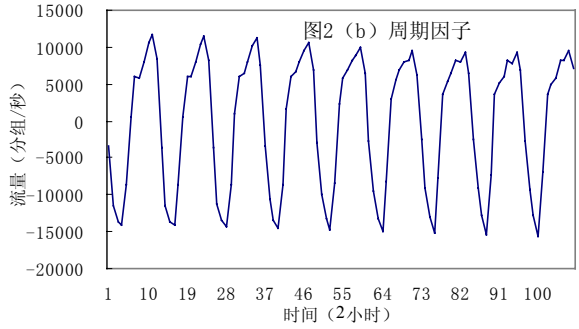
4. 实例研究

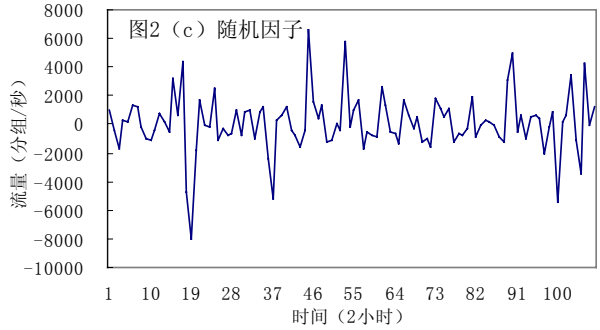
利用加法模型对图 1 中的流量资料进行详细分析。图 1 中的抽样时间尺度为 2 小时，以天为周期，一天共抽样 12 次，样本数取 N=120，（留最后一天 12 个资料作预报检验）。

4.1 模型选择和加法调整结果

由图 1 可知测量时间内流量增长趋势不是很明显，但周期波动很明显，且周期波动不随时间的推移发生明显变化，由第 2 节可知该资料选用加法模型进行季节调整比较合适。$X_t=T_t+C_t+E_t$，根据第 3 节的 3 种加法算法，对$\{x_t\}$进行加法处理，分解得到得到序列$\{T_t\}$, $\{C_t\}$, $\{E_t\}$，分别见图 2 中（a）、（b）、（c）图所示。







4.2 预测

流量行为分解分析的一个重要用途是进行流量预测。利用流量序列$\{x_t\}$的分解结果，可对第 N 个周期的序列值 $x_{N\times S+1}, x_{N\times S+2}, \ldots, x_{(N+1)\times S}$ 进行预报，S 为序列周期（本例 S=12），预报公式表示为：

$$\hat{x}_{N\times S+l}=\hat{T}_{N\times S+l}+\hat{C}_{N\times S+l} \quad （6）$$

其中：（1） $\hat{C}_{N\times S+l}=C_{(N-1)\times S+l}+C_{(N-2)\times S+l}$

（2）$\hat{T}_{N\times S+l}$可根据 ARIMA（2，1，0）模型 $(1+1.6461B-0.8864B^2)(1-B)x_t=a_t$ 外推。为了比较分解模型预测结果，使用 ARIMA 季节模型[7]直接对该流量时间序列进行一个周期的预测。通过检测，用于该实例使用 ARIMA（0，1，2）×（0，1，1）12 较为符合要求，其参数估计为表 1。使用预报误差 error 比较两种模型效果。

表 1：ARIMA 季节模型参数估计表

参数	估计值	标准差
非季节 MA,1 阶	0.2701	0.0913
非季节 MA,2 阶	0.3087	0.0917
季节 MA，12 阶	0.8811	0.078

$$error=\sqrt{\frac{\sum_{i=n+1}^{n+1+r}(X_i-\hat{X}_i)^2}{r}} \quad （7）$$

式（7）中，n 为序列中用于建模的时间长度，r 为预测的长度。cernet trace 中 n＝108，r 为 12。模型 error 统计量比较结果见表 2。累加模型通过简单的加法运算，利用分解出的趋势项和周期项进行预测，同 ARIMA 相比不需要复杂模型选择和参数估计。另外，预测时，如果考虑随机项的预测子成分，累加模型的预测精度会有所提高。

表 2：模型 error 统计量

模型	Cernet error
分解模型	6.61
ARIMA 模型	6.76

5. 结束语

由于大规模网络是复杂的非线性大系统，其流量行为也是复杂多变，因此目前国内

外对流量行为的研究还没有成熟的方法。本文根据非线性网络流量的特点，提出将复杂的流量分解成各个简单的子系统，通过研究各子系统行为来获得对复杂的流量总体行为认识。为此，提出一种简单的加法模型，通过不同的加法运算实现趋势项、周期项和随机项的分解。并进行流量预测研究，分别对趋势项、周期项的预测以实现流量总体行为的预测。本文加法模型的基础是假设流量时间序列图中存在天为周期，对于流量时间序列中存在的隐含周期需要通过相应的周期分析才能得到，这是将来需要进行的工作。

参考文献

[1] Kevin Thompson, Gregory J. Miller, and Rick Wilder, Wide-Area Internet Traffic Patterns and Characteristics (Extended Version), IEEE Network, November/December 1997.

[2] W. E. Leland, M. S. Taqqu, W. Willinger, D. V. Wilson, On the Self-Similar Nature of Ethernet Traffic, IEEE/ACM Transaction on Networking, Feb. 1994, vol.2, PP.1-15.

[3] V. Paxson, S. Flod, Wide-area traffic: The failure of possion modeling, Proceedings of the ACM/SIGCOMM'94, 1994, PP. 257-268.

[4] Rich Wolski, Forecasting Network Performance to Support Dynamic Scheduling Using the Network Weather Service, http://www-cse.ucsd.edu/users/rich/, 2001.5.

[5] S. Basu and A. Mukherjee. Time series models for internet traffic. Technical Report GIT-CC-95-27, Georgia Institure of Technology, 1996.

[6] M. E. Crovella, A. Bestavros, Self-similarity in world wide web traffic: Evidence and possible causes, IEEE/ACM Tranctions on Networking, vol. 6, Dec. 1997.

[7] 杨位钦、顾岚，时间序列分析与动态资料建模，北京理工大学出版社，1988.

Addition Decomposed Research of Traffic Behavior in a Large-Scale Network

Cheng Guang　Gong Jian　Ding Wei

(Computer Department of Southeast University　Nanjing 210096)

Abstract: Traffic behavior in a large-scale network is very perplexing and can be viewed as a complicated non-linear system. So far the research on traffic behavior doesn't have a well-rounded method. In this paper, according to the character of non-linear network traffic, the traffic behavior is decomposed into trend item, period item, and random item by a addition decomposed model. With such a decomposition, complicated traffic can be simulated by the compound of the three simpler sub-series with different mathematical tools. In order to check our model, the long-term traffic behavior of the CERNET backbone network is analyzed using the decomposed model, and the results are compared with ARIMA model. According to prediction error function value, We find that the model has the advantage of simplicity and high precision to describe the traffic marco-behavior.

Keywords: Additional Decomposed Model, Traffic Behavior, non-linearity, Prediction

基金项目：本课题受“863-317-01-03-99”课题资助

作者简介：程光，男，73.2，博士生，研究方向网络管理。龚俭，男，57.8，教授、博士生导师，研究方向网络安全、网络管理、网络体系结构。