大规模网络流量宏观行为周期性分析研究

程 光　龚 俭
（东南大学计算机系　南京　210096）

摘要：由于影响大规模网络流量行为长期变化的因素比较多，使得宏观流量周期行为属于非严格的周期，是统计与模糊概念上的周期。本文根据大规模网络宏观流量序列可能存在多日周期特性，提出了一个模糊方差周期分析模型，使得宏观流量过程周期的分析建立在严格概率统计的基础上。由于宏观流量时间序列具有明显的趋势行为，因此在进行周期分析之前，提出了先采用灰色系统理论从原始流量时间序列中分离出趋势项。对经过 CERNET 国家主干路由器长达 4 个多月流量序列周期分析表明，CERNET 的网络流量宏观行为具有 7 天周期规律的性质。

关键词：周期、灰色理论、模糊、假设检验
分类号： TP393　　文献标识码：A

1. 引言

Internet 作为由上亿台计算机互联而成的全球数据网络，最近几年一直在持续快速地膨胀发展，Internet 行为的研究也因此成为一项非常具有挑战意义的工作。尽管 Internet 的设计一直在不断地完善，但人们对网络行为许多方面的理解却较少，Internet 技术和管理的多样性、网络规模持续增长性、及其应用和使用方式的变化特性，都对网络行为研究提出了挑战，从而使 Internet 行为学研究从网络管理中分离出来，成为一门独立的网络研究科学。大规模 IP 网络是一个复杂的非线性系统，在其中运行的流量行为往往复杂多变，数据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，因而通过研究网络流量行为来理解网络行为相对困难。网络流量行为的描述是网络行为学的重要研究内容，它是网络规划、设计和管理的基本依据，对网络运行的 QoS 保证和网络安全也有很大影响。

目前国内外对流量行为研究比较活跃。1994 年 Nancy 和 George[1]利用 ARIMA(p,d,q)×(P,D,Q)s 季节模型对 NSFNET 主干网络流量进行了预测,1996 年 Sabyasachi[2]等人对时间序列模型用于 Internet 流量预测进行了理论分析，同时建立校园网和以太网的预测模型。1997 年 Rich[3]等人将 ARIMA 模型用于网络气象服务。94 年 Leland[4]等人对从 1989 到 1992 年测量的以太网数据进行分析首次发现了以太网 LAN 流量具有自相似性质。其后，95 年 V. Paxson 和 S. Floyd[5]分析了 89 年到 95 年 24 种不同的广域网数据发现 TCP 分组级到达符合自相似规律。1995 年波士顿大学的 Mark E. Crovella 和 Azer Bestavros[6]对一个广域网的 WWW 记录文件研究发现 WWW 流量具有自相似性。天津大学张连芳[7]等人分析了自相似流量序列的 FARIMA 模型。由此可见，流量行为的研究已引起众多科研工作者的重视并也取得了一定的成果。但目前流量行为分析主要停留在微观时间尺度领域，宏观流量行为研究主要的工作只是停留在使用常用的时间序列工具 ARIMA 进行流量预测研究，而对于宏观流量行为的运行规律和本质的研究工作很少。

大时间尺度的流量宏观行为在不同的时间粒度下有很大的变化与差异，主要原因是由于网络用户行为和网络故障的随机性造成的。由于从宏观上看网络用户行为存在一定的周期性变化，从而导致网络流量行为可能存在某些周期。1994 年 Nancy 和 George[1]对长达近 5 年的 NSFNET 主干网络流量记录文件进行分析，发现流量序列具有以年为周期的性质。2000 年东南大学网络中心在对 CERNET 华东(北)地区网的研究中发现流量序列具有以日为周期的性质[8]。由于影响网络流量行为长期变化的因素比较多，因此在日和年周期行为之间，网络流量还可能存在其它的非严格周期行为，这些周期是统计与模糊概念上的周期，不能象日周期和年周期行为那样可以直观地从流量时间序列图中直接得出。本文根据宏观流量序列可能存在多日周期的特性，应用灰色系统模型和模糊方差分析方法，对 CERNET 国家主干路由器上长达 4 个多月流量过程的宏观周期行为进行了严格概率统计基础上的分析。由于宏观流量时间序列具有明显的趋势行为，因此在进行周期分析之前，先采用灰色系统理

论从流量时间序列中分割出趋势项。下面先讨论建立灰色系统模型进行流量趋势项分解，然后建立流量残差模糊周期分析模型，并对实际流量进行周期分析，同时使用流量序列的自相关函数图证明了本文模型的合理性和优越性，最后总结归纳全文。

2. 趋势分析

从灰色系统理论的观点看，随机变量可视为在一定范围内变化的灰色量，而随机过程可视为在一定范围内变化的灰色过程。灰色系统理论认为：对于一个时序系统，必然是具有已知的信息，又有未知或不确知信息，且处于被动变化中。尽管系统行为现象是朦胧的，数据是杂乱的，但它毕竟是有序的，在这杂乱无章的数据背后，必然潜在某种规律。

由于流量要素自身变化的复杂性受诸多因素制约，具有很大的不确定性，其实质上就是一个处于动态变化中的灰色系统，因此可用 GM（1，1）[9]建模。

2.1 时序方程

GM(1,1)模型是单序列的一阶线性动态模型，分析下列流量序列：

$$x^{(0)}=\{x_1^{(0)},x_2^{(0)},\cdots,x_n^{(0)}\} \qquad (1)$$

由它产生的 1 次累加生成 1－AGO 为：

$$x^{(1)}=\{x_1^{(1)},x_2^{(1)},\cdots,x_n^{(1)}\} \qquad (2)$$

式中 $x_t^{(1)}=\sum_{i=1}^{t}x_i^{(0)}=x_{t-1}^{(1)}+x_t^{(0)}$ （t=1,2,…,n）

由于生成序列接近指数曲线，故认为是光滑离散系数，可用微分方程描述。一阶带系数的微分方程表达式为：

$$\frac{dx^{(1)}}{dt}+ax^{(1)}=b \qquad (3)$$

其解（即时序方程）可写成

$$x^{(1)}(t+1)=(x^{(1)}(0)-\frac{b}{a})e^{-at}+\frac{b}{a} \qquad (4)$$

式中参数 a、b 为待估参数。

2.2 参数估计

参数估计应用最小二乘法来求解，写成矩阵形式：

$Y=XB$

其中

$$X=\begin{bmatrix}-\frac{1}{2}\left|x^{(1)}(1)+x^{(1)}(2)\right| & 1\\ -\frac{1}{2}\left|x^{(1)}(2)+x^{(1)}(3)\right| & 1\\ \vdots & \vdots\\ -\frac{1}{2}\left|x^{(1)}(n-1)+x^{(1)}(n)\right| & 1\end{bmatrix}\quad Y=\begin{bmatrix}x^{(0)}(2)\\ x^{(0)}(3)\\ \vdots\\ x^{(0)}(n)\end{bmatrix}\quad B=\begin{bmatrix}a\\ b\end{bmatrix} \qquad (5)$$

则：$B=(X'X)^{-1}X'Y$

2.3 还原计算

确定了参数 a、b 之后，可计算出 $\hat{x}(t)$，则其还原数据为：

$$\hat{x}^{(0)}(t)=x^{(1)}(t)-x^{(1)}(t-1) \qquad (6)$$

根据式（6）可以从原始流量时间序列中剔除趋势项，然后对吞吐量残差序列进行下面的模糊周期分析研究。

3. 周期分析

大规模网络宏观流量序列各个数据的差异，主要由随机因素与网络用户行为的系统条件变化两方面原因造成，由随机因素引起的差异是随机差异，由用户行为变化引起的差异为条件差异。因用户行为的周期性变化，使得宏观流量序列的周期性变化。由于影响宏观流量序列变化的因素十分复杂，这里所谓的周期只是概率意义上的周期，需要通过周期分析来确定。提取时间序列的隐含周期有多种方法，如谱分析、谐波分析、方差分析等。但对网络流量来说，以方差分析法最为简单、适用，是目前较为常用的方法。本文对日流量序列的周期分析采用方差分析法。

日流量时间序列的时序变化过程可以看成是有限个不同的周期波叠加的结果，其数学模型为

$$x(t)=p_1(t)+p_2(t)+\cdots+p_n(t)=\sum_{i=1}^{n}p_i(t) \qquad (7)$$

式中　x(t)为日流量时间序列；$p_i(t)$为各个周期波的序列。

但直接从日流量序列的外形难以判断它是否存在某种周期，故需要先试验周期的天数为 b，b 应在 2，3，---，（n-1）/2（n 为奇数）或 n/2（n 为偶数）日中逐一试验取定。

某日流量序列 x_i, i=1,2,…,n。设该径流序列存在 b 日周期，将此 n 个数据按 b 日的时间间隔分组，得到 b 组不同数据，各组数据间的离差平方和为

$$S_1=\sum_{j=1}^{b}a_j(\bar{x}_j-\bar{x})^2 \qquad (8)$$

组内离差平方和为

$$S_2=\sum_{j=1}^{b}\sum_{i=1}^{a_j}(x_{ij}-\bar{x}_j)^2 \qquad (9)$$

a_j 为第 j 组数据的项数；x_{ij} 为第 j 组数据的第 i 个数值；$\bar{x}=\frac{1}{n}\sum_{j=1}^{b}\sum_{i=1}^{a_j}x_{ij}$ 为日流量序列 x_i 的平均值，

$\bar{x}_j=\frac{1}{a_j}\sum_{i=1}^{a_j}x_{ij}$ 为第 j 组平均值。令 $f_1=b-1$，$f_2=n-b$，设各组数据相互独立，且均服从方差相同的正态分布，则统计量 $F=\left(\frac{S_1}{f_1}\right)\Big/\left(\frac{S_2}{f_2}\right)$ 服从 $F(f_1,f_2)$分布。

设给定一个显著性水平区间[α_1，α_2]，$\alpha_1>\alpha_2$，定义非严格周期 b 日的相对隶属函数为：

$$\mu(F)=\begin{cases}1 & ,F\geq F_{\alpha_1}(f_1,f_2)\\ \dfrac{F-F_{\alpha_2}(f_1,f_2)}{F_{\alpha_1}(f_1,f_2)-F_{\alpha_2}(f_1,f_2)} & ,F_{\alpha_2}(f_1,f_2)<F<F_{\alpha_1}(f_1,f_2)\\ 0 & ,F\leq F_{\alpha_2}(f_1,f_2)\end{cases} \qquad (10)$$

$F_{\alpha_1}(f_1,f_2)$，$F_{\alpha_2}(f_1,f_2)$ 分别为显著性水平 α 1，α 2，由 F 分布表查得的临界值，μ (F)表示在给定的显著性水平区间[α 1， α 2]条件下，存在非严格周期 b 日的相对隶属程度，$0 \le \mu(F) \le 1$。μ (F)=1 表示存在 b 日周期，μ (F)=0 表示不存在 b 日周期。模型（10）称为模糊假设检验模型。

特殊地，如果 α =α 1=α 2，即将显著性水平区间[α 1, α 2]变为点 α，则为通常的统计假设检验模型

$$\chi(F)=\begin{cases}1, & F \ge F_{\alpha}(f_1,f_2)\\ 0, & F < F_{\alpha}(f_1,f_2)\end{cases} \qquad (11)$$

χ（F）为特征函数，χ（F）=1 表示存在 b 日周期，χ（F）=0，则不存在 b 日周期。可以看出，式（11）是式（10）的一个特例。式（10）系将存在 b 日周期的问题，定义一个模糊集合，而式（11）则以普通集合论为基础。显然，用式（10）来表达日流量序列是否存在 b 日周期，比式（11）更符合现象的实际情况，这是本文提出的模糊假设检验模型的一个优点。

根据式（9）求得的统计量 F，如有 $F \ge F_{\alpha_1}(f_1,f_2)$，认为存在 b 日周期；如果 $F \le F_{\alpha_2}(f_1,f_2)$，认为不存在 b 日周期；若 $F < F_{\alpha_1}(f_1,f_2)$ 且 $F > F_{\alpha_2}(f_1,f_2)$，则可按式（4）中间一个等式求得存在 b 日周期的隶属度 μ (F)，然后再根据下面论述的原则来判断是否存在 b 日周期。

4. 流量分析

我们使用本文的模型对 2001 年上半年通过 CERNET 国家主干路由器四个月的流量序列进行分析。网络流量长期周期基本特性可用自相关函数 ACF(i)来反映，我们使用 Bartlett 估计法[10]进行自相关估计。图 1 为实测数据的自相关函数图，图 2 为剔除趋势项以后剩余流量序列的自相关函数图。图 3 为将趋势项、周期项剔除后的自相关函数图。

试验周期的天数b为2、3、…、66。现以b=7为例进行分析论述。本例中样本数为n=123，故f_1=b-1=6，f_2=n-b=117。根据日流量残差序列，由式（10）求得F=19.066。选定显著性水平区间：α 1 = 0.01，α 2 = 0.1。由 F 分布表查得 $F_{\alpha_1}(f_1,f_2)=F_{0.01}(6,117)=2.96$，$F_{\alpha_2}(f_1,f_2)=F_{0.1}(6,117)=1.83$。则有 $F > F_{\alpha_1}(f_1,f_2)$，根据式（10）有 μ （F）=1，存在7日周期性，其显著性水平为0.01，或判断正确的保证率为99%。在一

图1：实测流量自相关函数

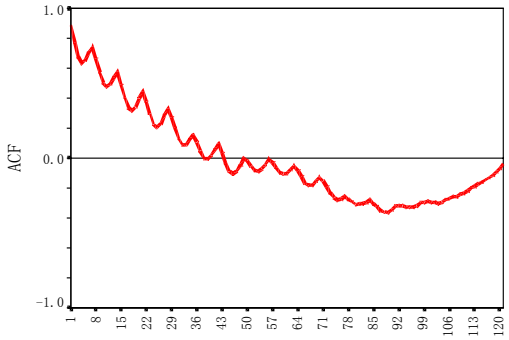


图2：剔除趋势项后序列的自相关函数

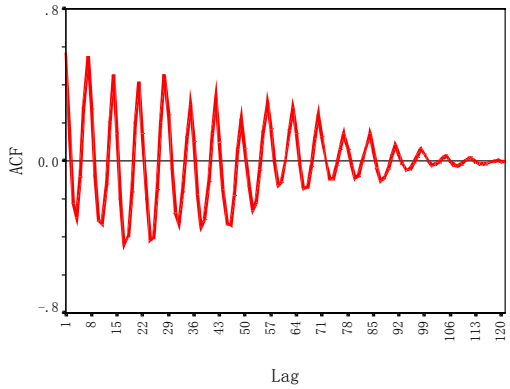
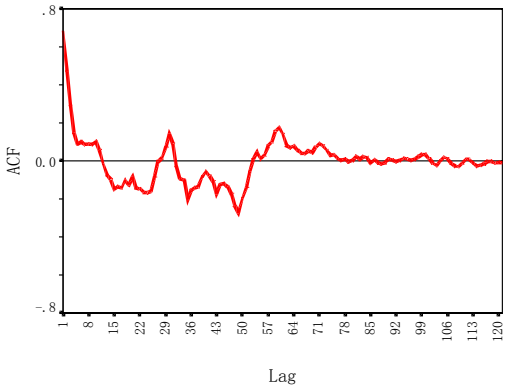


图3：剔除趋势项、周期项后的序列自相关函数



个周期内，峰顶在星期六和星期日之间波动，而峰谷在星期二和星期三之间波动，峰顶和峰谷之间相隔3天到4天。网络用户日周期行为的语义背景是比较清楚的，它反映了用户对上网时间的选择。而这种行为周期所反映的用户行为语义则较为复杂，它一定程度上反映了CERNET用户对网络的依赖性和使用的后效性。Thompson等人[11]在MCI商业主干网络两个不同测试点所测得的流量时序图，也存在7天为周期的特性，这显然和商业网络用户的生活行为以7日为周期有关。但在他们的时序图中，峰顶位于星期二到星期四之间，峰谷位于星期六和星期天之间，这个结果同我们研究结果正好相反，这种区别表明以学生为主体的用户群和以商业用户为主体的用户群在上网行为上有很大不同。当然，对于流量周期行为产生原因的更深入研究需要进一步分析流量中不同网络应用组成成分及会话数等统计量的变化等因素，这也是我们将来要继续进行的工作。

类似地，可以计算不同试验周期 b 的统计量，表 1 给出 b＝2～10 的结果，可见通过 CERNET 国家主干路由器流量序列至少存在以 7 日为周期的周期特性。同时表 1 中还给出了未剔除趋势项对原始数据直接进行周期检验的统计量 UF，及其对应的 μ (UF)。

表 1　b=2~10 试验周期的成果表

b	f_1	f_2	$F_{0.01}$	$F_{0.10}$	F	μ (F)	是否存在 b 日周期	UF	μ (UF)
2	1	121	6.85	2.75	0.483	0	不存在	0.10	0
3	2	120	4.79	2.35	0.258	0	不存在	0.10	0
4	3	119	3.95	2.13	0.161	0	不存在	0.04	0
5	4	118	3.48	1.99	0.274	0	不存在	0.04	0
6	5	117	3.18	1.90	0.366	0	不存在	0.10	0
7	6	116	2.96	1.83	19.07	1	存在	2.28	0.40
8	7	115	2.80	1.77	0.134	0	不存在	0.03	0
9	8	114	2.67	1.73	0.348	0	不存在	0.11	0
10	9	113	2.56	1.69	0.774	0	不存在	0.26	0

从表 1 可以知道，直接对流量序列进行周期分析的效果没有剔除趋势项后再进行周期分析效果显著，对其原因，我们可以直接从流量序列自相关函数图可以知道。图 1 是实测数据自相关函数图，没有剔除趋势项，由于图 1 中趋势项在总流量序列成分中起决定优势，将周期项掩盖了，因此从图 1 中很难看出流量序列中存在什么周期行为。图 2 是将趋势项剔除后剩余序列自相关函数图，由于剔除了趋势项，残余流量序列中周期项起了决定因素，因此，图 2 中可以明显看出流量序列存在以 7 天为周期的行为。图 3 是剔除趋势项、7 日为周期的周期项后剩余随机项的自相关函数图，图中可以看出，当 lag 增加时，ACF 趋进于 0，这种被负指数控制的衰减形式，其图象如一条越来越小的尾巴，这种行为符合 AR（p）自回归模型的性质，AR（p）模型是一种平稳序列，其中不含周期项和趋势项。流量序列自相关函数图的分析结果同我们模型的分析结果完全相吻合，证明了我们模型的合理性和优越性。

5. 结论

本文提出了大规模网络宏观流量序列周期分析的模糊统计模型。模型考虑到流量序列中存在趋势行为，提出了使用灰色系统理论分解趋势项，研究结果表明，剔除趋势项后的周期分析比直接对原始数据进行周期分析效果显著。在进行周期性分析时，既考虑到流量序列的周期系统变化所引起的条件差异，也考虑由于随机因素所引起的随机差异，据此判定流量序列是否存在周期，比目前常用的假设检验模型，理论上更为严谨，实用上可以避免或减少判定的差错。如果把显著性水平区间：$[\alpha_1, \alpha_2]$缩为一点值 α，则文中提出的模糊假设检验模型就变为通常的假设检验。本文还使用流量序列自相关函数图证明了本文模型的合理性和优越性。

本文对 2001 年上半年通过 CERNET 国家主干路由器 4 个多月的日平均吞吐量的研究结果表明：CERNET 网络流量的 7 日周期规律比较明显。大规模网络流量的这种周期性是网络用户行为随星期变动有密切的联系，并不是一种偶然的现象。由于国家法定节日和寒暑假的影响以及一些不确定因素

的影响，用户随星期周期变动的行为又具有模糊性和不确定性，因而不能象日周期和年周期可以直接从时间序列图中看出，本文的模糊周期分析模型体现了流量这一特点。虽然我们分析的只是 CERNET 网络流量长期行为规律，对于其它网络流量的周期行为不一定符合 7 日周期的行为，但同样可以用本文的模型进行分析。

参 考 文 献

[1] N. Groschwitz, G. Polyzos, A Time Series Model of Long-term Traffic on the NSFnet Backbone, In Proceedings of the IEEE International Conference on Communications(ICC'94), May 1994, PP. 164 –168.

[2] S. Basu and A. Mukherjee. Time series models for internet traffic. INFOCOM '96. Fifteenth Annual Joint Conference of the IEEE Computer Societies. Networking the Next Generation. Proceedings IEEE , Volume: 2 ,1996, PP: 611 -620.

[3] Rich Wolski, Forecasting Network Performance to Support Dynamic Scheduling Using the Network Weather Service. http://www-cse.ucsd.edu/users/rich, 2000.11.

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

[5] V. Paxson, S. Flod, Wide-area traffic: The failure of possion modeling, IEEE/ACM Transactions on Networking, vol.3, June 1995, PP. 226-244.

[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., pp: 160-169

[7] 张连芳, 薛飞, 王雷等, 自相似网络业务的一个 FARIMA 模型, 计算机研究与发展, Vol. 37, No 9, 2000. 9. pp: 1138-1144.

[8] Hua Wu, Jian Gong, Forecast of Network Behavior Based on Singular-Spectrum Analysis, In: Proceedings of the 16th IFIP World Computer Congress (wcc2000), August, 2000, Beijing, China. PP: 1230-1236.

[9] 邓聚龙, "灰色预测与决策", 华中工学院出版社, 武汉, 1986.8.

[10] 杨位钦, 顾岚, 时间序列分析与动态数据建模, 北京理工大学出版社, 北京, 1988.

[11] Kevin Thompson, Gregory J. Miller, and Rick Wilder, Wide-Area Internet Traffic Patterns and Characteristics (Extended Version), IEEE Network, November/December 1997, pp: 234-242.

The Periodicity Research on Traffic Macro-Behavior in a Large-Scale Network

Cheng Guang Gong Jian

（the Computer Department of Southeast University NanJing 210096）

Abstract： The facts that influences the long-term changing of traffic behavior in a large-scale network are very complicate, which makes the periodicity of traffic macro-behavior not be a strict period, rather than a periodicity in the statistical and fuzzy concept. In the paper, according to the periodicity that the macro-traffic time series may possess, we put forward a periodical analysis model based on fuzzy variance. The model makes the traffic macro-period analysis based on the strict statistical theory. Because the macro-traffic time series have obviously trend behavior, so at first the gray system theory is used to decompose the trend item of original macro-traffic time series, then the periodicity of traffic time series is analyzed. Based on the real traffic measured from the CERNET backbone router for four months in 2001, it can be showed that the traffic macro-behavior in a large-scale network has the character of seven-day period.

Keywords: period, gray theory, fuzzy, hypothesis verification

基金项目：本课题受“863-317-01-33-99”课题资助

作者简介：程光，男，28 岁，博士研究生，主要研究方向为网络行为学、网络管理。龚俭，男，44 岁，教授、博士生导师，主要研究方向为网络安全、网络管理、网络体系结构。