基于小波的网络流量分解模型

程 光, 龚 俭, 丁 伟

(东南大学 计算机科学与工程系, 江苏 南京 210096)
(江苏省计算机网络技术重点实验室, 江苏 南京 210096)
E-mail: gcheng@njnet.edu.cn

摘 要: 大规模网络的流量行为体现为一个相当复杂的非线性系统, 目前国内外对它的研究还没有成熟的方法. 多分辨小波分析能将交织在一起的不同频率成分组成的复杂时间序列分解成频率不相同的子序列. 基于小波分解和重构思想, 文章将流量过程分解成不同尺度下的小波系数和尺度系数, 然后分别重构最高层低频序列和各层的高频序列. 从各子频率序列中分离趋势项、周期项和随机项, 并对各成分分别分析与建模, 最后合成原流量时间序列的流量预测模型. 通过CERNET流量实例分析表明该模型的精度高于ARMA模型.

关 键 词: 小波; 时间序列; 网络流量; 分解
中图分类号: TP393 **文献标识码**: A **文 章 编 号**: 1000-1220(2005)03-0400-04

Traffic Decomposed Model on Wavelet in a Large-Scale Network

CHENG Guang, GONG Jian, DING Wei

(*Department of Computer Science & Engineering, Southeast University, Nanjing* 210096, *China*)
(*Key Lab of Computer Network Technology Jiangsu Province, Nanjing* 210096, *China*)

Abstract: Traffic behavior in a large-scale network is very perplexing, so far the research on traffic behavior doesn't have a well-rounded method. By multi-resolution analysis, the complex traffic time series can be decomposed into many different frequent components. In the paper, based on the wavelet decomposed and recomposed theory, the backbone network traffic are decomposed wavelet coefficients and scale coefficients of different scales. Then the top layer low frequency and all layers high frequency time series are recomposed. And the trend term, period term and random term can be decomposed from these recomposed frequency series, so every sub-series can be analyzed and modeled separately. Finally, the traffic forecasting model can be built by recomposed the sub-series models. The model is proved through CERNET traffic and its precision is larger than ARMA model.

Key words: wavelet; time series; traffic analysis; decompose

1 引 言

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. 由于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, 所以在描述网络流量行为时误差较大. 如: AR[1](Auto Regressive)模型、MA(Moving Average)模型和ARMA[2](Auto Regressive Moving Average)模型用于解决平稳过程, ARIMA[2](Auto Regressive Integrated Moving Average)模型和ARIMA季节模型[3, 4]用于处理齐次的非平稳性过程等. 这些模型都是假设网络流量序列是平稳的, 或通过差分法以分离趋势成分、周期成分和随机成分来进行流量分析和建模, 这些方法的主要缺点是丢失流量序列中重要的周期信息和趋势信息. 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, 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, 而这些特点难以用传统模型来描述.

论文使用多分辨小波分析方法对网络流量进行分析, 使用Mallat算法将原始流量一层层分解成不重迭的子信号, 由于分解后的信号在频率上比较单一, 可以对每个子成分单独进行重构和建模, 分别进行分析和预测, 最后再合成得到原时间序列的预测值. Aimin Sang[5]使用ARIMA模型证明网络流量的子序列分解模型的预测精度和可预测性均不小于直接对原始序列进行预测.

2 流量分解模型

流量时间序列X(t), 一般由趋势项A(t)、周期项P(t)和随机项R(t)组成, 它们通常包含在不同时间尺度的子序列中. 对X(t)经过若干次小波分解, 将其分解成不同尺度成分.

收稿日期: 2003-09-03 基金项目: 国家自然科学基金重点项目(90104031)资助; 国家"九七三"项目(2003CB314803)资助 作者简介: 程光, 男, 1973年生, 博士, 研究方向为网络测量和网络行为学; 龚俭, 男, 1957年生, 教授, 博导, 研究方向为网络安全, 网络体系结构; 丁伟, 女, 1962年生, 教授, 研究方向为网络管理, 网络测量

实际上趋势项为大尺度成分, 因此子序列中只有$A_n^d f$ 包含趋势成分, 其它每个子序列成分中包含周期项和随机项 通过这种思路, 可以将一个复杂的非线性流量时间序列问题分解成多个尺度、多种组成成分的流量子序列, 对于每个子序列可以用相关的数学模型来描述, 使复杂的问题分解成多个简单的问题

2.1　多分辨分析尺度分解

多分辨分析是对低频部分进一步分解, 而对高频部分则不予考虑 分解具有关系: $X = D_n + D_{n-1} + \cdots + D_1 + A_n$, 其中 $D_i (i \in [1, n])$ 是信号 X 分解成的高频部分, A_n 是低频部分, 如果要进行进一步分解, 还可以将低频部分 A_n 分解成低频部分 A_{n+1} 和高频部分 D_{n+1}, 以下再分解依次类推 设 V_j 表示分解中的低频部分 A_j, W_j 表示分解中的高频部分 D_j, 则 W_j 是 V_j 在 V_{j-1} 中的正交补, 即 $V_j \oplus W_j = V_{j-1}$, 则多分辨分析的子空间 V_0 可以用有限个子空间来逼近, 即

$$V_0 = V_1 \oplus W_1 = V_2 \oplus W_2 \oplus W_1 = \cdots = V_n \oplus W_n \oplus W_{n-1} \oplus \cdots \oplus W_2 \oplus W_1 \tag{1}$$

如 $f_j \in V_j$ 代表分辨率为 2^{-j} 的序列 $f \in L^2(R)$ 的逼近, 而 $d_j \in W_j$ 代表逼近的误差, 则 (1) 式意味

$$f_0 = f_1 + f_d = f_2 + d_2 + d_1 = \cdots = f_n + d_n + d_{n-1} + \cdots + d_2 + d_1 \tag{2}$$

因此, 任何序列 $f \in L^2(R)$ 可以根据分辨率 2^{-n} 时 f 的低频部分和分辨率 2^{-j} 下的高频部分完全重构 小波多分辨分析可用 Mallat[6] 算法

2.2　趋势项分解

经小波分解后的子序列, 只有 $A_n^d f$ 中包含有趋势项, 可以使用多项式拟合公式

$$A_n^d f = a_0 + a_1 t + a_2 t^2 + \cdots + a_k t^k \tag{3}$$

当 $k = 0$, 趋势项为均值函数, 即 $A_n^d f = a_0$, 此时为常值趋势, 当 $k = 1$ 时, $A_n^d f = a_0 + a_1 t$, 即趋势为随时间线性变化

2.3　周期项分解

已分解的 $A_n^d f$ 序列和 $D_i f$ 序列包含有周期成分和随机成分, 周期成分使用傅立叶级数法分解 设 $D_i f$ 序列中包含 n 个点, $n = 2l + 1$. 因此 $D_i f$ 可以近似成有限项傅立叶级数, 即:

$$D_i f = \frac{a_0}{2} + \sum_{i=1}^{l} (a_i \cos i\bar{\omega}t + b_i \sin i\bar{\omega}t) + \epsilon(t) \tag{4}$$

式中, $\epsilon(t)$ 为随机成分, $\omega = 2\pi/n = 2\pi f$, f 表示 n 点的范围仅表现为一个周期 a_0、a_j 及 b_j 为系数, 即:

$$\begin{cases} a_0 = \overline{D_i f} \\ a_i = \dfrac{2}{n} \sum_{i=1}^{n} D_i f(t) \cos i\bar{\omega}t \\ b_i = \dfrac{2}{n} \sum_{i=1}^{n} D_i f(t) \sin i\bar{\omega}t \end{cases} \tag{5}$$

$$|c_i| = \frac{1}{2}\sqrt{a_i^2 + b_i^2} \tag{6}$$

$|c_i|$ 为振幅, 当振幅最大时, 其对应的频率为主频率, 继而取出第二、第三个频率, 将这些周期对应的各分量 $(a_i \cos i\bar{\omega}t + b_i \sin i\bar{\omega}t)$ 相叠加, 即为所求周期项的结果

2.4　随机项分解

将各子序列分离出周期项后, 剩余的就是随机项, 随机项可以近似为平稳随机项和纯随机项, 可以使用 AR(p) 模型分解平稳时间序列项

$$x(t) = \beta_{p,1} x(t-1) + \beta_{p,2} x(t-2) + \cdots + \beta_{p,p} x(t-p) \tag{7}$$

式中: $\beta_{p,j}$ $(j = 1, 2, \cdots, p)$ 为自回归系数, p 为模型阶数 算法描述如下:

(1) 计算模型系数 自回归系数利用最小二乘法, 建立 Yule-Walker 方程组, 采用下列递推公式求解:

$$\begin{cases} \beta_{1,1} = \gamma_1 \\ \beta_{k,k} = \dfrac{\gamma_k - \sum_{j=1}^{k-1} \beta_{k-1,k} \gamma_{k-j}}{1 - \sum_{j=1}^{k-1} \beta_{k-1,j} \gamma_j} \quad (k = 2, 3, \cdots) \\ \beta_{k,j} = \beta_{k-1,j} - \beta_{k,k} \beta_{k-1,k-j} \quad (j = 1, 2, \cdots, k-1) \end{cases} \tag{8}$$

式中, $\beta_{\cdot,j}$ 为自回归系数, γ_k 为 X3(t) 的 k 阶样本自相关系数:

$$\gamma_k = \frac{\sum_{t=1}^{n-k} X3_t X3_{t+k}}{\sum_{t=1}^{n} X3_t^2} \tag{9}$$

(2) 计算模型阶数 可通过 AIC 准则来确定:

$$AIC = \min\left\{ n \ln \frac{\sum (X3(t) - \overline{X3})^2}{n - P - 1} \right\} \tag{10}$$

3　流量分析模型

使用 Mallat 算法对流量序列 $f_0(t)$ 分解并重构成各子成分, 如式 (2), 一般考虑采用三层分解 式中 f_n 表示第 n 层低频成分 (趋势) 重构, d_i 表示第 i 层的高频重构, 假设 f_n 可以用模型 $\hat{f}(t)$ 仿真, d_i 可用 $\hat{d}_i(t)$ 来仿真, 因此对于流量序列的第 k 步预测模型可以表示

$$\hat{f}_0(t+k) = \hat{f}_n(t+k) + \sum_{i=1}^{n} \hat{d}_i(t+k) \tag{11}$$

下面分别对 f_n, d_i 建仿真模型

(1) 提取 f_n 的趋势成分 $Tf_n(t)$, 将趋势成分从 f_n 中减去, $df_n = f_n - Tf_n$;

(2) 从 df_n 和 d_i 中提取周期成分为 Pdf_n 和 Pd_i, 相应分离周期成分为:

$$ddf_n = df_n - Pdf_n;\ dd_i = d_i - Pd_i \tag{12}$$

(3) 分别对 ddf_n 和 dd_i 使用 AR(p) 模型建模, 分离平稳随机成分, 相应的模型分别为 $rddf_n$ 和 rdd_i,

(4) 建立仿真模型

$$\begin{aligned} \hat{f}_n(t) &= Tf_n(t) + Pdf_n(t) + rddf_n(t) \\ \hat{d}_i(t) &= Pd_i(t) + rdd_i(t) \end{aligned} \tag{13}$$

根据 (18)、(20) 由此可以得出 $f_0(t)$ 的预测模型为

$$\hat{f}_0(t) = Tf_n(t) + Pdf_n(t) + rddf_n(t) + \sum_{i=1}^{n}(Pd_i(t) + rdd_i(t)) \tag{14}$$

4 流量实例分析

为了验证上述模型, 论文用的网络流量数据来自于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 CERNET 华东(北)地区网络是CERNET 全国8 个地区网络之一, 连接江苏、安徽、山东的150 所高等院校和研究单位 2000 年下半年当该地区网与CERNET 主干网的互联信道从OC- 3 升级至OC- 48 时, 当天两网之间的流量高峰由原来的12000 个分组/秒迅速上升到35000 分组/秒, 每天高峰流量和低峰流量的比值也由原来的1. 5 倍增加到4 倍(cernet, 时间粒度为日, 见图1). 首先是将分解模型分别作用于上述三组数据, 求出各自的

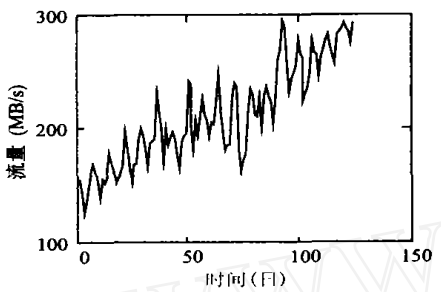

图1 CERNET 序列图

参数, 然后是基于相关性分析的特征分析和分解模型与传统的ARMA 季节模型的比较结果 在流量序列分解和重构中, 小波函数的选取是相当重要的, 不同的小波函数应用的性能具有很大差别, 需要根据具体的应用来决定小波函数的选取 在这里我们选择db3 小波函数

流量时间序列图见图1, 小波分解三层图形见图2, 各层

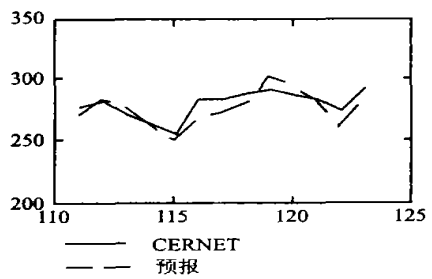

图2 预报序列比较图

分解系数重构见图3 所示, 使用前110 天数据对数据进行预测, 将后13 天的预测结果同实测流量相比较见图2 下面使用ARMA (7, 0, 0) × (0, 1, 0) 7: 参数 (β, β_2, …, β_7) = (0. 6606, 0. 1631, -0. 0805, -0. 1232, -0. 0085, 0. 1721, -0. 2153) 建立预报模型 同时定义预报误差error (15) 式比较两种模型效果

$$error = \sqrt{\frac{\sum_{i=n+1+r}^{n+1+r}(X_i - \hat{X}_i)^2}{r}} \tag{15}$$

式中, 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 为110, r 为13. 表1 为cernet 的多分辨分析模型和ARMA 模型预报的error 统计量 从表1 可知, 多分辨分析模型描述流量模型更为精确

表1 模型error 统计量表

模 型	error 统计量
多分辨分析模型	67. 83
ARMA 季节模型	123. 56

从图3 和图4 可以看出, 趋势成分集中在低频, 而高频仅

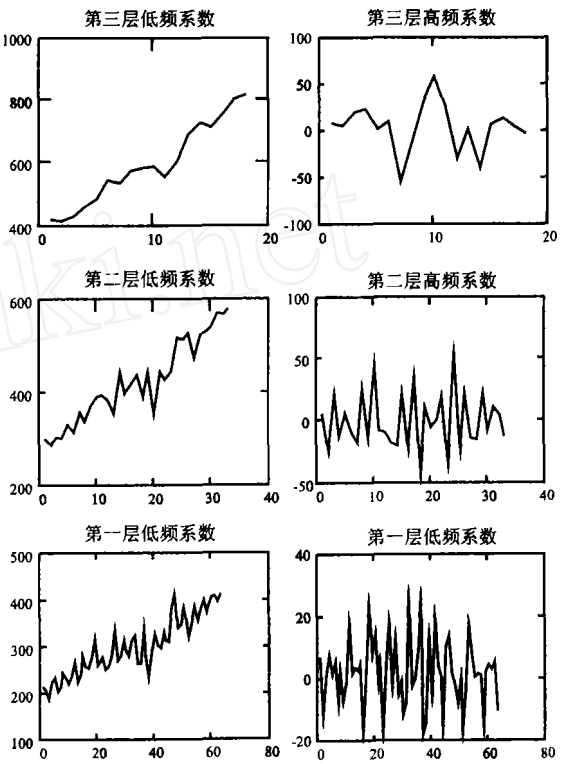

图3 时间序列三层分解频率系数

包含周期项和随机项, 且如果继续将低频进一步分解, 可以使最终分解结果仅含有可以用多项式拟合的趋势项 超高频由于振幅不大, 且频率很大, 可以认为仅含有随机项, 在周期项中可以不作考虑 因此第三层低频重构需要考虑趋势项、周期项和随机项, 第二层高频和第三层高频需要考虑周期项和随机项, 对于第一层高频可以只考虑随机项或将其看成是纯随机项作为误差考虑

5 结 论

大规模网络流量序列中由于各种因素的影响, 其行为表现为非平稳时间序列, 用常用的ARMA 等平稳线性模型难以估计和建模 而多分辨分析的最大优点是可以将复杂的流

量时间序列按不同的尺度分解成不同的层次, 这使得流量序列变得简单. 由于分解子序列将趋势项、周期项和随机项变得容易分解, 因此可以使用多项式拟合公式、傅立叶级数、AR(p)自回归模型进一步将同时将各子序列进一步分解成各个子成分建模. 而模型中忽略的成分只有不可预测的纯随机成分.

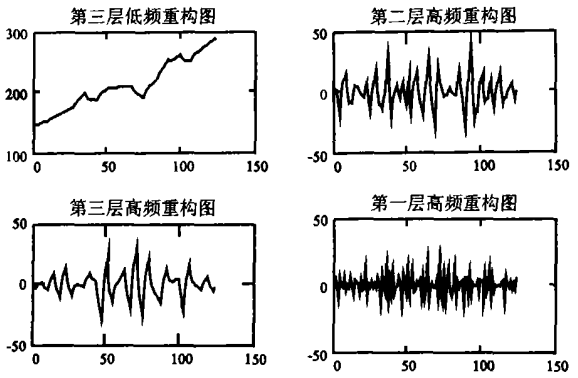

图 4　各层分解系数重构图

将多分辨分析模型同ARMA 季节模型的实验结果比较发现其精度高于ARMA 模型一倍左右. 分析其原因发现多分辨分析模型具有以下优点:

(1) 多分辨分析模型分别各个频率角度、各频率振幅以及从各组成成分描述流量行为, 同时使用更多的参数描述流量行为, 因此可以更准确更完整的描述流量行为规律.

(2) 多分辨分析使流量中的重要信息可以分解细一些, 而对于次要信息分解粗一些, 因此可以具有重点地描述流量行为.

多分辨分析在网络流量分析中已经具有一定的应用, 如 Rudolf H. Riedi 等人[7]将多分辨分析小波模型用于网络流量长相关的特性研究中. 小波是研究网络流量行为学有力工具, 将来需要做的工作是进一步将小波理论应用于微观流量行为和宏观流量行为分析中.

References:

[1] Rich Wolski. Forecasting network performance to support dynamic scheduling using the netwrk weather service[C]. In: Proceedings of the 6th IEEE Symposium on High Performonce Distributed Computing, 316-325, Los Alamitos, California, 1997, 8.

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

[3] Claffy K, Polyzos G C, Braun H W. Traffic characteristics of the t1 nsfnet backbone[C]. Proceedings IEEE INFOCOM '93, 885-892, March 28- April 1, 1993.

[4] Groschwitz N, Polyzos G. A time series model of long-term traffic on the NSFnet backbone[C]. In Proceedings of the IEEE International Conference on Communications (ICC '94), May 1994.

[5] Aimin Sang; San-qi Li, A predictability analysis of network traffic[C]. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, 2000, Volume: 1, 342-351 vol. 1.

[6] Liu Gui-zhong, Di Shuang-liang, Analysis and application of wavelet[M]. Xidian University the Press, 1992.

[7] Rudolf H R, Matthew S C, Vinay J R, Richard G B. A multifractal wavelet model with application to network traffic[J]. IEEE Transactions on Information Theory, Apr. 1999, 45(3): 992-1018.

附中文参考文献:

[6] 刘贵忠, 邸双亮. 小波分析及其应用[M]. 西安: 西安电子科技大学出版社, 1992.