

基于小波多尺度的网络论坛话题热度趋势预测

张虹¹, 赵兵², 钟华¹

(1. 北京城市学院 人工智能研究所, 北京 100083;

2. 中国电力科学研究院, 北京 100085)

摘要:文中基于小波多尺度分析进行网络论坛话题热度趋势的预报。该方法主要是对由帖子的点击数(或回复数)所形成的原始时间序列进行小波分解与重构,得到一个低频信号和多个不同尺度的高频信号;对具有近似平稳特征的低频信号建立 ARIMA 预测模型;对变化较多的各高频信号分别建立神经网络预测模型;然后分别对各信号进行一步预测并组合预测结果,获得网络论坛话题热度的最终预测。实验表明:将本方法用于网络论坛话题的热度趋势预测,可得出良好的预测精度。

关键词:时间序列;小波分解与重构;ARIMA 模型;神经网络

中图分类号:TP311

文献标识码:A

文章编号:1673-629X(2009)04-0076-04

Hot Trend Prediction of Network Forum Topic Based on Wavelet Multi-Resolution Analysis

ZHANG Hong¹, ZHAO Bing², ZHONG Hua¹

(1. Artificial Intelligence Institute, Beijing City University, Beijing 100083, China;

2. China Electric Power Research Institute, Beijing 100085, China)

Abstract: In this paper, a method was proposed, which combined wavelet multi-resolution analysis to forecast network forum topic hot trend. That was first, using wavelet multi-resolution to decompose and reconstruct the original time series formed from click numbers or reply numbers of forum post, to create one low frequency signal and several high frequency signals. Then, the approximate low frequency signal is predicted using ARIMA model and the high frequency signals are forecasted respectively using neural networks that have different parameters. After one-step-ahead prediction, the predicted results of these signals are combined into the final predicted result of network forum topic hot trend. This proposed method was tested to make better prediction accuracy for network forum topic hot trend.

Key words: time series; wavelet decomposition and reconstruction; ARIMA model; neural network

0 引言

随着 Internet 网络的普及和互联网用户的增多,以文本信息为载体的网络论坛(BBS)已经成为人们获取信息、发表言论的重要场所。同时,基于用户行为分析的网络论坛管理也成为广泛关注的研究方向。网络论坛话题的形成与发展过程就是在互动性的作用下,由少数人向多数人传播的过程,即高级用户(意见领袖)牵头→中级用户反馈→普通用户点击、浏览。高级用户作为论坛主体的代表,发帖数量较大,影响力较强,其发帖的受关注程度(如点击数和回复数)很高,即

所谓的热帖。这一传播过程启示我们,如果能从帖子点击数(或回复数)所形成的时间序列进行网络论坛话题的热度趋势预测,则可以根据预测挖掘潜在的高级用户,进而通过对高级用户的分析实现对网络论坛的管理和分析。

由于网络论坛本身是复杂的非线性系统,每个帖子受关注程度(如点击数和回复数)所形成的时间序列往往复杂多变,又呈现非线性升、降趋势,而这些特点难以用传统的时间序列模型描述^[1]。而传统的应用范围较广的 ARIMA(自回归求和滑动平均)时间序列分析方法虽然适合处理非平稳时间序列,但是在网络论坛话题热度预测的过程中,点击数(或回复数)所形成的时间序列受随机干扰因素影响大,而 ARIMA 模型的参数是固定的,不能很好地适应不确定性强的要求。近年来,神经网络、隐马尔科夫链等预测算法被相继提

收稿日期:2008-07-23

基金项目:国家高技术研究发展计划(863 计划)研究资助项目(2005AA147030)

作者简介:张虹(1972-),女,山西太原人,讲师,从事 Web 数据挖掘、模式识别研究。

出^[2],使预测精度得到提高。而小波分析方法利用其多分辨分析的特点可以提取时间序列的频谱特性,在时频两域都具有表征信号局部特征的能力。

文中利用小波变换对非平稳时间序列进行小波分解与单支重构,分解后的各信号频率成分更加单纯;然后,根据各分解信号的不同特点,对较为平稳的低频信号采用 ARIMA 模型预测,对变化较多的高频信号采用神经网络预测;最后将各信号预测值加以合成,获得最终预测值。实验结果表明,本方法用于网络论坛话题的热度趋势预测,显示出较好的预测精度。

1 方法和模型

1.1 小波分解与重构

小波变换具有多分辨分析(MRA)的特点,Mallat^[3]在 MRA 的基础上给出了小波系数快速分解的金字塔算法,并构造了用于小波分解和重构的高、低通滤波器组,大大简化了小波系数的计算。要先确定小波函数以及小波分解的低通滤波器 H ,高通滤波器 G 及其对偶算子 H^* 、 G^* ,选择分解层数 L 。

定义 令 X 表示预测原始时间序列, $X = \{x_1, x_2, \dots, x_N\}$, 其中 x_i 为第 i 个采样点的实际值, N 为时间序列的长度。对原始序列进行小波分解的过程为:

$$X_{i+1} = HX_i; Y_{i+1} = GX_i; i = 0, 1, 2, \dots, L \quad (1)$$

其中, X_i 和 Y_i 分别是分辨率为 2^{-i} 的原始信号 X 的逼近信号(低频信号)和细节信号(高频信号)。

当分解进行到第 L 层时,得到第 1 至 L 层共 L 个高频信号和 1 个第 L 层低频信号: $Y_j = \{y_{j,i}\}, j = 1, 2, \dots, L; X_L = \{x_{L,i}\}$, 其中, Y_j 表示第 j 层高频信号, X_L 表示第 L 层低频信号。

由于每进行一层分解所得的高频信号和低频信号的点数都会比分解前的信号减少一倍,这对预测是不利的,为了使各分支的长度保持不变,对各分支进行单支重构,对这 $L+1$ 个信号利用 Mallat 重构算法:

$$X_{i-1} = H^* X_i + G^* Y_i, i = L, L-1, \dots, 1 \quad (2)$$

重构后的 $L+1$ 个分支与原始序列的长度一致,即: $\tilde{Y}_j = \{\tilde{y}_{j,i}, 1 \leq i \leq N\}, j = 1, 2, \dots, L; \tilde{X}_L = \{\tilde{x}_{L,i}, 1 \leq i \leq N\}$, 其中 \tilde{Y}_j 为重构后的第 j 层高频信号, \tilde{X}_L 为重构后的第 L 层低频信号。

低频信号和各高频信号变换特点各不相同。低频信号比较平滑,较好地体现了时间序列的周期性,采用 ARIMA 模型预测即可。各高频信号则变化较多:低频信号非线性较强,是短时间依赖关系的体现;高层信号非线性较弱,是长时间依赖关系的体现^[4,5]。基于高频信号的这种规律特性,分别建立不同参数的 BP 神

经网络进行预测。

1.2 低频信号的 ARIMA 模型预测

ARIMA 模型^[6]预测的基本思路是:对于非平稳的时间序列,用若干次差分(称之为“求和”)使其成为平稳序列,用 ARMA(p, q)模型对该平稳序列建模,之后经反变换得到原序列。

用数学公式表示这样一个 ARIMA(p, d, q)过程如下:

$$\varphi(B) \nabla^d x_t = \theta(B) a_t$$

其中, x_t 和 a_t 分别表示原序列和白噪声序列。

$$\varphi(B) = 1 - \varphi_1 B - \dots - \varphi_p B^p = 1 - \sum_{k=1}^p \varphi_k B^k \neq 0 \quad (3)$$

$$\theta(B) = 1 - \theta_1 B - \dots - \theta_q B^q = 1 - \sum_{k=1}^q \theta_k B^k \neq 0$$

$\nabla = 1 - B$ 为差分算子, B 为后移算子,即 $Bx_t = x_{t-1}$ 。

$\nabla^d = (1 - B)^d$ 是 d 阶差分。 $d = 1$, 进行一次差分处理,即 $\nabla w_t = w_t - w_{t-1}$; $d = 2$, 进行两次差分处理,即 $\nabla^2 w_t = \nabla w_t - \nabla w_{t-1}$, 依此类推。

ARIMA 模型的建立分为 3 个阶段:

(1)模型识别,根据既定数据利用自相关系数和偏自相关系数来判断应该使用哪个模型,阶数为多少。

(2)估计选定模型中的参数。

(3)检验模型是否合理(残差的随机性分析)并预测。ARIMA 模型不仅计算复杂,而且模型的识别困难。有时可能会给出几种不同的模型。有鉴于此, Akaike 和 Ozaki 提出了确定模型阶数的信息准则,简称 AIC 准则(Akaike Information Criterion)。该准则规定 AIC 值最小的模型为最优化模型。本研究即利用这一准则来建立模型,利用确定出的最佳模型预测即得低频信号的预测值 $\hat{x}_{L,N+1}$ 。

1.3 高频信号的 BP 神经网络模型预测

文中对 L 个细节信号分别建立 BP 神经网络预测模型^[7],每个模型均为输入层、隐层和输出层三层。因为只作单步预测,各 BP 模型的输出层单元只有一个。由于进行了单支重构,每层分解信号的输入层单元数不变,隐层单元数大约取输入单元数的一半。将每一个高频信号时间序列按比例划分为两组,分别作为 BP 模型训练数据和测试数据。应用训练后的模型进行一步预测,可得预测值 $\hat{y}_{1,N+1}, \hat{y}_{2,N+1}, \dots, \hat{y}_{L,N+1}$ 。

1.4 预测值组合算法

预测值的合成方法有很多,文中采用最简单的方式,将各分支预测值直接对应相加,即:

$$\hat{x}_{N+1} = \hat{y}_{1,N+1} + \hat{y}_{2,N+1} + \dots + \hat{y}_{L,N+1} + \hat{x}_{L,N+1} \quad (4)$$

2 仿真实例

本次实验选取了网易新闻论坛 2005 年 05 月 18 日到 2006 年 01 月 22 日共 600 条帖子的点击数所形成的时间序列作为实验对象,采集间隔为 1 小时,每条帖子给出了 132 个数据点,将所收集的数据分为两组,前 5.5 天的数据作为训练集,后 1 天的数据作为测试集。预测采用单步滚动方式。

对原始点击数时间序列选用 DB4 小波,将数据分解到第 3 层。图 1 为点击数时间序列的三层小波分解和重构。其中 S 是原始信号, A3 是第 3 层低频信号, D1、D2 和 D3 分别是第 1 至 3 层高频信号。

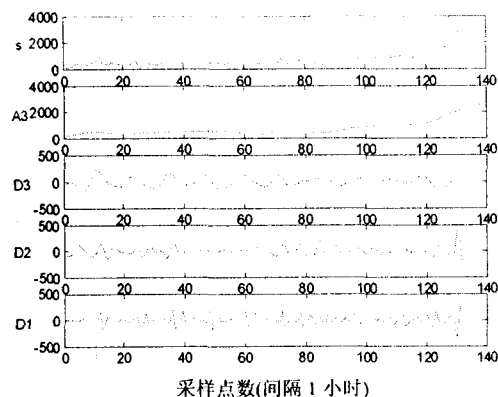


图 1 点击数时间序列的分解和重构

从低频信号的序列图看出,该序列明显是非平稳的。为使序列符合运用 ARIMA 模型的条件,先对序列进行自然对数转换以平稳序列的方差,并对序列进行一阶逐期差分和一阶季节差分,以消除序列的趋势和周期性。对原序列的一阶逐期差分和一阶季节差分的自相关函数估计值和偏自相关函数估计值的变化趋势分析,发现序列前后的方差波动已平稳,周期性也被消除,得到一组平稳的随机序列。据此,可初步确定模型应是 $ARIMA(p,1,q)(P,1,Q)^s$ 。通过观察差分后序列的 ACF 和 PACF 图,并根据经验初步判断模型为 $ARIMA(2,1,1)(1,1,1)^s$ 或 $ARIM(2,1,0)(1,1,0)^s$,然后根据模型的拟合优度、残差情况以及系数间的相关性估计模型参数并对模型进行检验,如此反复进行

检验和调试,直至得到最佳模型^[8]。

模型的参数估计见表 1,拟合优度统计见表 2。结果显示 $ARLMA(2,1,0)(1,1,0)^s$ 模型所有参数都有统计学意义;拟合优度统计量有标准误差、对数似然函数值、Akaike 信息准则 (AIC)、Schwarz 贝叶斯准则 (SBC),数据显示 $ARIMA(2,1,0)(1,1,0)^s$ 拟合优度最好; $ARIMA(2,1,0)(1,1,0)^s$ 模型参数无明显相关性;对残差序列作自相关图,结果显示 $ARIMA(2,1,0)(1,1,0)^s$ 模型的 BoxLjung 统计量均无统计学意义 ($P > 0.05$)。可以认为,所选模型 $ARIMA(2,1,0)(1,1,0)^s$ 是恰当的。

表 2 模型的拟合优度统计量对比

	$ARIMA(2,1,0)(1,1,0)^s$	$ARIMA(2,1,1)(1,1,1)^s$
Number of Residuals	132	132
Standard error	0.260	0.273
Log likelihood	-8.972	-9.131
AIC	23.944	28.261
SBC	31.962	41.626

这里将 3 层高频信号的训练集分别送入 BP 神经网络进行训练,网络结构为 $20 \times 10 \times 1$ 。

图 2 为帖子点击数预测值与实际值的比较。

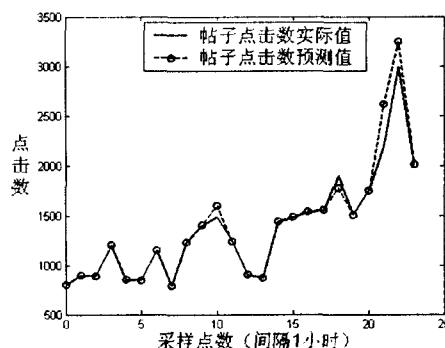


图 2 帖子点击数预测值与实际值的比较

为评估本算法的性能,基于同一组数据建立 $ARI-MA(4,1,1)(1,1,0)^s$ 模型,进行单步预测,并计算其均方误差式(5),结果见表 3。

表 1 模型参数估计结果

$ARIMA(2,1,0)(1,1,0)^s$					$ARIMA(2,1,1)(1,1,1)^s$				
	Estimates	Std Error	t	Approx Sig		Estimates	Std Error	t	Approx Sig
AR1	-0.745	0.090	-8.286	0.000	AR1	-0.605	0.228	-2.649	0.009
AR2	-0.339	0.091	-3.734	0.000	AR2	-0.304	0.145	-2.092	0.039
SAR1	-0.426	0.100	-4.274	0.000	MA1	0.126	0.241	0.525	0.601
					SAR1	0.333	0.303	1.101	0.273
					SMA1	0.970	2.710	0.358	0.721

表 3 两种预测方法的均方误差

预测方法	文中方法	ARIMA(4,1,1)(1,1,0) [*] 模型
均方误差	7.501	13.476

$$\sigma = \sum_{i=1}^N \frac{(X_i - \hat{X}_i)^2}{N} \tag{5}$$

其中, X_i 是真实值, \hat{X}_i 是预测值, N 是样本数。

显然,文中方法明显优于 ARIMA 预测方法。

3 结束语

文中以网络论坛话题热度趋势预报为研究对象,针对小波分解后各信号不同特点建立不同预测模型,仿真结果表明,此方法用来预测网络论坛话题热度趋势这一类非平稳的时间序列取得了较好的预测效果。但是,在实验中发现:小波函数的选择是影响模型预测精度的关键因素之一,应对该问题作进一步研究;ARIMA 的应用前提是时间序列的平稳性,实际工作中数据往往是非平稳序列,需对序列进行预处理,使之达到平稳的要求。

未来的工作主要在于根据预测值将那些普通帖子过滤掉,充分考虑热帖内容,挖掘潜在的高级用户,发现更有意义的高级用户行为模式,进而实现对网络论坛的管理和分析。

(上接第 75 页)

一定测距精度的基础上,文中方法所需的计算时间比较少,与其它两种方法相比,更加能够满足实际雷达系统实时性的要求。

4 结束语

针对 FMCE 雷达单目标距离谱的特点,提出一种新的精度改进算法,可以在一定程度上提高 FMCW 雷达的测距精度。该算法是基于 FFT 变换,在距离谱谱峰位置对目标距离所对应的频率进行估值运算。文中的方法与直接 FFT 变换法,FFT - CZT 法,内插采样法相比较有如下优点:计算精度更高,计算时间更少即计算量更小。

此外,雷达测距精度还受到调频非线性^[6]、噪声^[7]等因素的影响,要想获得更高的测距精度,必须对调频非线性进行校正,对回波信号进行滤波降噪处理。

参考文献:

[1] 陈祝明,丁义元,向敬成.采用 Chirp - Z 变换提高 LFM-

参考文献:

[1] Hansen J V, Nelson R D. Neural networks and traditional time series methods: a synergistic combination in state economic forecasts[J]. IEEE Trans Neural Networks, 1997, 8(4): 863 - 873.
[2] Khotanzad A, Sadek N. Multi - scale high - speed network traffic prediction using combination of neural networks[C]// Proceedings of the International Joint Conference on Neural Networks. Portland, O R, USA: IEEE Press, 2003: 1071 - 1075.
[3] Mallat S. A Theory of Multi - resolution Signal Decomposition: The Wavelet Transform[J]. IEEE Trans on PAMI, 1989, 11(7): 674 - 693.
[4] 孙明谦,姚淑萍,胡昌振.服务器负载的小波 - 神经网络 - ARMA 预测[J]. 计算机工程与应用, 2007, 43(10): 154 - 155.
[5] 冉启文,单永正,王 琪,等.电力系统短期负荷预测的小波 - 神经网络 - PARMA 方法[J]. 中国电机工程学报, 2003, 23(3): 38 - 42.
[6] 常学将,陈 敏,王明生.时间序列分析[M]. 北京:高等教育出版社, 1993.
[7] 周 波,石爱国,蔡 烽,等.基于多尺度分析和神经网络的时间序列预报[J]. 计算机应用与软件, 2005, 22(1): 93 - 94.
[8] 薛 薇.SPSS 统计分析方法及应用[M]. 北京:电子工业出版社, 2004.

CW 雷达的测距离精度[J]. 信号处理, 2002, 18(2): 110 - 112.
[2] Zhang Jie. Research of High Precision Frequency Measure Algorithm Based on LabVIEW[C]//2007. ICEMI '07. 8th International Conference on Electronic Measurement and Instruments. Xi'an, China: [s. n.], 2007.
[3] 李 政,张容权,杨建宇,等.利用频域增采样内插方法提高 LFM CW 雷达测距精度[J]. 电讯技术, 2005(5): 77 - 80.
[4] 刘 宝,刘军民.FMCW 雷达快速高精度测距算法[J]. 电子测量与仪器学报, 2001(9): 41 - 45.
[5] 宋 玮.FMCW 雷达测距精度及其信号处理技术的研究[D]. 南京:南京理工大学, 2004.
[6] Ahmed N. Hardware and Software Techniques to Linearize the Frequency Sweep of FMCW Radar for Range Resolution Improvement[D]. B. S. E. E. : University of Kansas, 2004.
[7] Krzysztof K S. Novel Method of Decreasing Influence of Phase Noise on FMCW Radar[C]//Proceedings of 2001 CIE International Conference on Radar. Beijing, China: [s. n.], 2001: 319 - 323.