基于流量特征的网络流量预测研究

张凤荔 赵永亮 王 丹 王 豪
(电子科技大学计算机科学与工程学院 成都 611731)

摘 要 传统的非线性模型已经不再适用于网络流量建模,为了能够更精确地对网络流量建模,必须考虑到网络流量的特性。针对网络流量的自相似、长度分布、周期等特征进行分析,结合小波变换与时间序列模型,有效地建立流量预测模型。首先对流量的自相似和平稳性进行分析,并对长度、周期等特征进行描述,其次根据实际流量的自相似性和平稳性选择小波变换与时间序列相结合的方法进行建模,产生预测结果,最后根据长度与周期特征粗略判断预测的合理性。根据实验验证与分析,该方法具有极大的灵活性,相比单一的小波-FARIMA模型可以减少大量的运算,同时能够描述网络流量的短相关与长相关特性。

关键词 流量特征,小波变换,流量预测
中图法分类号 TP393.01 **文献标识码** A

Prediction of Network Traffic Based on Traffic Characteristics

ZHANG Feng-li ZHAO Yong-liang WANG Dan WANG Hao
(School of Computer Science & Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

Abstract The traditional model such as nonlinear model could not adapt to the model of network traffic. So only considering these characteristics can researchers model the network traffic accurately. By combining the analyses on self-similarity, length distribution and period of network traffic, making use of the wavelet transform and time series model to predict the traffic, and finally, comparing the length distribution and periodic, we can know whether the prediction result is reasonable. Firstly, the characteristics of network traffic such as self-similar and stationary were analyzed. Secondly, based on the result of the first step, the model was constructed and prediction results were obtained through selecting wavelet transform and time series. Finally, taking advantages of the length distribution and period, the model's flexibility and accuracy were verfied. Through some experiments, it is proved that our model can reduce some computing compared with w-farima model and reflect the short-dependence and long-dependence of network traffic.

Keywords Traffic characteristics, Wavelet transform, Traffic prediction

1 引言

网络流量已经被验证具有自相似性[1]、长相关、多分形等特性,传统的基于泊松的模型与线性时间序列模型已经不能有效地对网络流量序列进行拟合。因此,许多学者利用小波分析、神经网络、支持向量机、混沌理论等模型或方法对网络流量建模,以期能够较好地反映出网络流量的特性。如文献[2]使用自适应信号分解方法对原始信号进行分解,然后分别采用神经网络和ARIMA模型进行预测,虽然这种预测方法提高了准确性,但神经网络的一些缺陷仍难以避免。文献[3]提出一种紧致型小波神经网络流量预测算法,采用了真实校园网流量数据进行实验,并与相关方法对比,也取得了有效的成果。但该方法同时采用了小波与神经网络相结合的方式,增加了复杂度。文献[4]结合了混沌理论和支持向量机,构造了混合模型用于网络流量的短期预测,它能够较大程度地用于描述短相关性,对于长相关性则是力有未逮。文献[5]对移动网络的流量序列进行分析,在分析流量周期性的基础上使用乘积季节自回归综合移动平均模型对流量序列进行预测。该方法适用于季节性比较强的流量序列,对于短期和规律性不强的网络流量的预测效果不是很好。文献[6]引入了支持向量机作为流量建模的基本模型,但支持向量机对大规模的文本支持不足,一般要与其他方法相结合才能较好地应用。文献[7]中作者利用小波对信号的处理能力来对网络流量进行建模,用LMK代替了LMS来改进算法,减少了大量的计算。文献[8]则是使用模糊理论对网络流量进行了分析,但如何控制模糊理论预测的结果使其不容易超出范围也是一个比较重要的问题。

小波理论的方法可以消除网络流量的长相关性,而时间序列模型的方法比较适用于平稳性较好的流量序列[9]。基于这两种方法的优点,提出基于网络流量的自相似、长度分布等

到稿日期:2013-06-25 返修日期:2013-11-15 本文受国家自然科学基金(61133016),工信部科技重大专项(2011ZX03002-002-03)资助。
张凤荔(1963—),女,博士,教授,博士生导师,主要研究方向为网络安全、移动数据管理及其应用等,E-mail:fzhang@uestc.edu.cn;赵永亮(1990—),男,硕士,主要研究方向为网络安全、网络行为分析、流量分析等;王 丹(1987—),女,硕士,主要研究方向为网络安全;王 豪(1988—),男,硕士,主要研究方向为网络安全。

特点与时间序列的平稳性，采用小波变换与时间序列模型（AR、ARMA、ARIMA、FARIMA）相结合的方式对网络流量进行建模，并使用流量的周期、长度等特征粗略验证预测的合理性。

本文第 2 节对网络流量特征长度、周期等进行了简单的分析。第 3 节给出了基于小波变换与时间序列模型的混合预测算法，至于自相似的计算分析则是采用传统的 R/S 图法进行计算，本文不再给出相关的内容介绍。第 4 节对实验的过程及实验的结果进行了详细的分析。最后则是对全文的总结以及下一步研究工作的展望。

2 网络流量特征研究

2.1 数据集描述

本文所用的实验数据集来自贝尔实验室所采集的流量数据[10]与 DataMarket 网站上所收集的时间序列数据集[11]，这些数据包括视频流量、局域网和广域网流量。本文主要使用 BC-pAu89. TL 和 United Kingdom 校园骨干网流量数据。其中，BC-pAu89 数据集则是以 1s 为时间间隔，选取 3124 条数据进行建模研究，并选取 1000 条作为特征研究；United Kingdom 则是选用以 1 小时为间隔的一个月的数据。图 1、图 2 分别是两个不同的数据集。

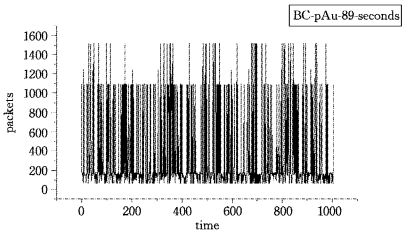

图 1 BC-pAu89 s 级数据

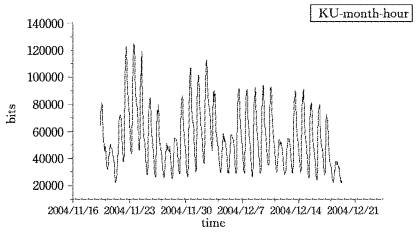

图 2 United Kingdom 数据集

2.2 流量长度、周期特征分析

从图 1、图 2 可以看出，流量具有明显的周期性与突发性，对于流量大小周期分布的研究可以以流量的突发性作为出发点，总结流量在一段时间内的大致分布。

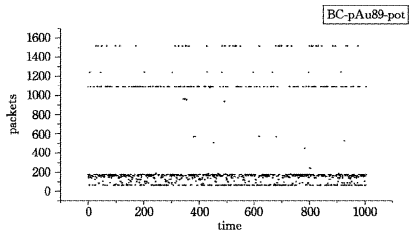

图 3 BC-pAu89 点状图

图 1 所示为 BC-pAu89 数据集，流量采集的间隔为 1s，其周期性不明显，周期的研究可以忽略不计，但从图 3 可以看到，在每秒的流量中 100～200 数据级的点数相对较多，1000～1200 相对较少，次之则是 1400 到 1600，因此进行预测时，以这 3 段数据范围进行划分，验证流量预测的结果时可以进行参考。

United Kingdom 的校园网流量是以小时作为时间间隔采集的，因此其分布更有规律一些。通过对图 2 所示的 30 天的数据分布进行分析，可以看出，流量具有明显的突发性与周期性，大概 2004-12-21 到 2004-12-24 是一个周期性的流量变化，下一阶段则从 2004-12-29 到 2005-12-04 重复类似的变化，而从 2004-12-24 到 2004-12-29 这段时间的流量也会发生突发性变化。其长度特征也随着周期性的变化而变化，这些周期性的数值与长度的变化都可以作为流量预测结果合理性的判断之一。

经过以上分析可以看出，对流量长度分布的研究如一段时间的数据流量大小的范围分布与大时间尺度下的周期分布都可以作为预测结果的合理性与在线实时预测的修正手段。

3 网络流量预测算法

3.1 小波变换与时间序列模型

3.1.1 小波变换

小波变换实际上是寻求标准正交基，然后将信号或者序列在这组正交基上进行分解，以便在分解的信号上能够做更合适的分析处理，最后重建原始尺度的信号。

本文使用的小波变换算法是 Mallat 算法[12]，它是在 1988 年由 S. Mallat 给出的正交小波基的构造方法以及正交小波变换的快速算法，其主要是构造两个带通滤波器，对样本进行频谱划分，使得信号样本被分解到不同的频率上。Mallat 算法的多尺度小波分解与合成的公式如下：

分解公式：

$$\left.\begin{aligned} c_{j+1}(n)&=\sum_{m\in z}c_j(m)h(m-2n)\\ d_{j+1}(n)&=\sum_{m\in z}c_j(m)g(m-2n)\end{aligned}\right\} \tag{1}$$

重构公式表示为：

$$c_j(n)=\sum_{m\in z}c_j(m)h(m-2n)+d_{j+1}(m)g(n-2m) \tag{2}$$

事实上，Mallat 算法的分解过程通过图示可以表示为一个倒金字塔形状，如图 4 所示。

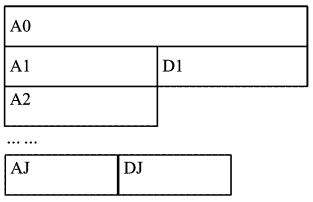

图 4 Mallat 算法分解过程

重构算法正好与分解算法相反，其图示是正向的金字塔形状，如图 5 所示。

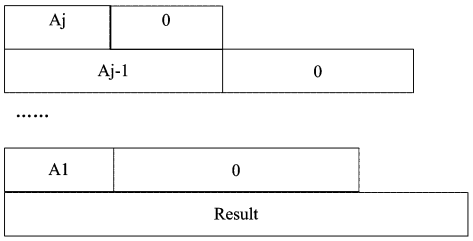

图 5 近似系数 Aj 的单支重构过程

3.1.2 时间序列模型

时间序列模型包括 AR 模型、ARMA 模型、ARIMA 模型

以及变体 FARIMA 模型，其中后两种模型都是由前面的 ARMA 模型变换而来的，而 FARIMA 模型又是 ARIMA 模型的差分系数取小数阶时得到的，因此这里只简要介绍 FARIMA 模型。

FARIMA 模型，又称分形自回归综合滑动平均模型，FARIMA(p,d,q)过程是 1980 年由 Hosking 提出的，是对 ARMA(p,q)过程的一种自然扩展，当 d 为 0 时即为 ARMA(p,q)模型，其定义如下。

$$\Phi(B)\nabla^d X_t=\theta(B)\varepsilon_t \tag{3}$$

其中，实数 $d\in(-0.5,0.5)$，且 $\Phi(B)$和 $\theta(B)$是 AR 和 MA 中的一部分，当 $d\in(0,0.5)$时，X_t 具有长相关特性的平稳可逆过程；当 $d=0$ 时，X_t 是短相关过程。

$$\nabla^d=(1-B)^d=\sum_k C_d^k(-B)^k \tag{4}$$

$$C_d^k(-1)^k=\Gamma(-d+k)/(\Gamma(-d)\Gamma(k+1)) \tag{5}$$

$\Gamma(\cdot)$表示 Γ 分布，其中 ε_t 为高斯分布。当 $k\to\infty$时，相关函数 $\rho(k)\sim\partial k^{2d-1}$，$\partial$为有限正值且与 k 无关。X 是渐进自相似的，且具有自相似参数 $H=d+1/2$。

FARIMA 模型的预测公式为：

$$\hat{X}(h)=\sum_{j=1}^{\infty}\pi_j^{(h)}\hat{X}_{t+h-j},$$

$$\pi_j^{(h)}=\pi_{j+h-1}-\sum_{i=1}^{h-1}\pi_i\pi_j^{h-1},\pi_j^{(1)}=\pi_j \tag{6}$$

系数向量 π 可由下式迭代得出：

$$\pi_j^{(1)}=\pi_j=\theta_1\pi_{j-1}+\theta_2\pi_{j-2}+\cdots+\theta_q\pi_{j-q}+\varphi_j,j>0 \tag{7}$$

其中，$\pi_0=-1$，当 $j>p+q$ 时，$\varphi_j=0$。

预测的均方误差定义为：

$$\hat{\sigma}_t^2(h)=E(X_{t+h}-\hat{X}_t(h))^2 \tag{8}$$

3.2 流量预测算法

算法思想：

该算法使用小波变换与时间序列相结合的方式来对实际流量进行建模分析，并得到预测结果。算法可以分为以下几步：

(1)检验流量序列的平稳性，如果序列是相对平稳的，则直接使用 ARMA 模型，否则使用 R/S 法估计实际流量的自相似参数 Hurst；如果 Hurst 参数值大于 1，并且流量序列也是不平稳的，则表明流量序列相关性不是很强，直接使用 ARIMA 模型进行预测分析；如果 Hurst 参数的值小于或者接近 0.5，则直接使用 ARMA 算法或者 ARIMA 算法进行预测，得到预测结果，转向第五步；如果 Hurst 参数值介于 0.5 至 1 之间，则进行下一步；

(2)使用小波分解 db3 小波对实际流量进行分解，使用 Mallat 算法进行单支重构，得到近似信号 A3 与细节信号 D1、D2、D3，并分别使用步骤(1)中的算法计算其 Hurst 值；

(3)如果 Hurst 参数值小于或接近 0.5 则使用步骤(1)中提到的 ARMA 或 ARIMA 算法；否则使用 FARIMA 算法进行预测；

(4)对步骤(3)中产生的结果使用合成算法也即 $Pred=D_3'+D_2'+D_1'+A_3'$(其中 D_3'、D_2'、D_1'、A_3'为各步预测结果)，组成最终结果；

(5)分析预测性能，计算预测误差等，并将第 2.2 节提到的方法与结果相结合，判断预测合理性。

具体的算法流程如图 6 所示。

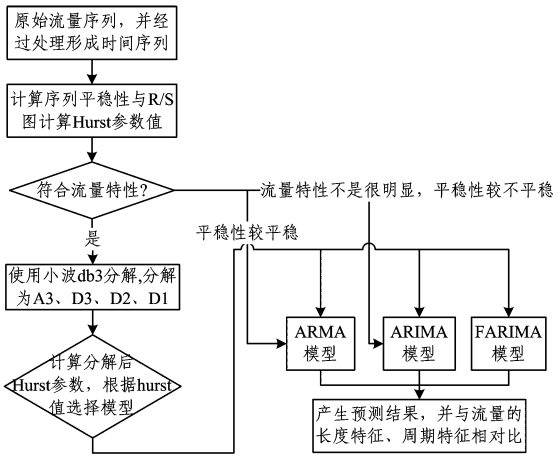

图 6　预测算法流程

4　实验及其结果分析

本文分别取 BC-pAu89、United Kingdom 校园网数据(等时间间隔 1s)3142 条、一周的数据(等时间间隔 1h)80 条作为样本，首先检验时间序列的平稳性，由平稳性检验结果得到 BC-pAu89 的序列是不平稳的，而 United Kingdom 的数据也是不平稳的，进一步计算 Hurst 参数值。

通过 R/S 图法估算可得，3142 条记录的 BC-pAu89 数据集的 Hurst 参数值为 0.6204，其 R/S 图 Hurst 估算如图 7 所示；而 United Kingdom 样本数据集的 Hurst 参数值大于 1，并且不是平稳的，根据算法直接使用 ARIMA 模型进行分析。分别对这两个数据集进行分析得到以下的结果。

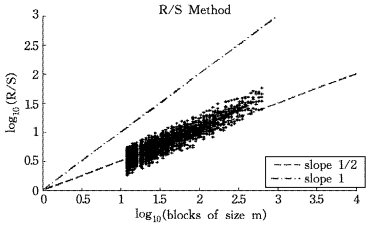

图 7　R/S 图估算 BC-pAu89 Hurst 值

通过以上的 Hurst 值计算以及 ADF 检验，可以看出 BC-pAu89 的 Hurst 参数值大于 0.5 而小于 1，因此使用 db3 小波分解其变量得到如下近似信号 CA3 与细节信号 CD1、CD2、CD3，并通过小波单支重构得到近似序列 A3 和细节序列 D3、D2、D1，如图 8 所示。

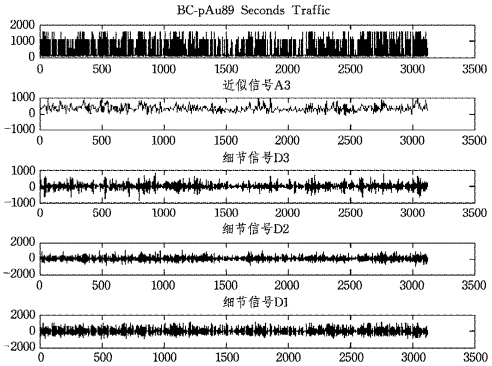

图 8　BC-pAu89 小波分解与单支重构的近似信号与细节信号

由于 A3 序列接近于原始序列，并且其 Hurst 值估算为 0.6931，因此使用 FARIMA 模型；D3 序列的 Hurst 值估算为 0.2397，相对平稳，使用 ARMA 模型预测即可；D2 序列的 Hurst 值估算为 0.1818，使用 ARMA 预测 D2 序列；D1 序列的 Hurst 值估计为 0.1073，尝试使用 ARMA 模型拟合。最终原始序列的预测值可以用 A3＋D3＋D2＋D1 的预测值小波重构得到。A3 信号如图 9(左)所示；而 D3 信号预测结果如图 9(右)所示。细节信号 CD2 和 CD1 的模型得到的预测结果分别为图 10(左)和(右)。

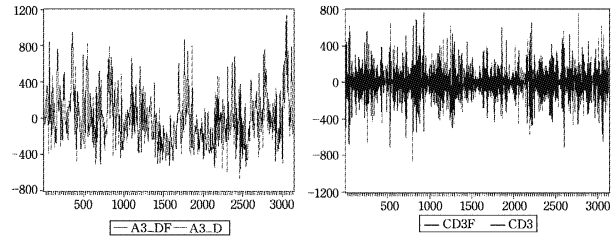

图 9　近似信号 A3 与细节信号 D3 预测拟合图

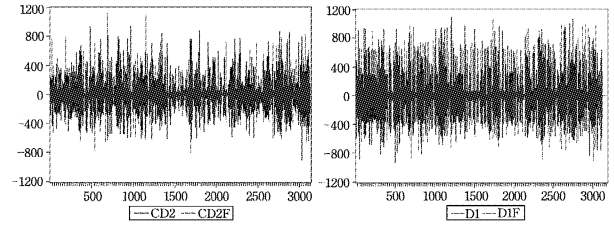

图 10　细节信号 D2 与 D1 预测拟合图

将预测信号 A3、D1、D2、D3 通过合成得到最终结果预测拟合图，如图 11 所示。

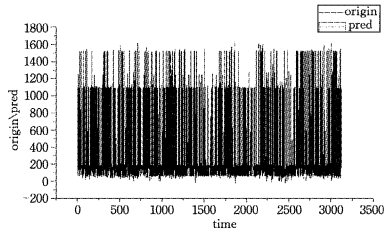

图 11　一周数据预测拟合图

而使用 FARIMA 模型拟合 BC-pAu89 数据集，得到的模型为 FARIMA(3,0.1204,2)，拟合结果如图 12 所示。

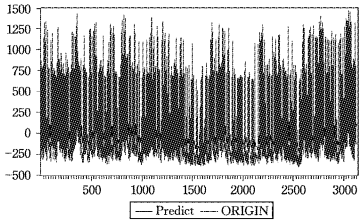

图 12　FARIMA 模型拟合 BC-pAu89 结果

对于 United Kingdom 样本数据集，由于其平稳性太差，而且 Hurst 参数值大于 1，因此需要进行整数阶差分，经过相关性检验和单元平方根检验，可知需要对 80 条小时级的样本数据集进行 2 阶差分，这样可得比较平稳的序列，进一步进行模型参数评估，得到最终的时间序列预测算法为 ARIMA(2,2,1)，最终的预测结果如图 13 所示，模型残差为 33.78，相对较高。

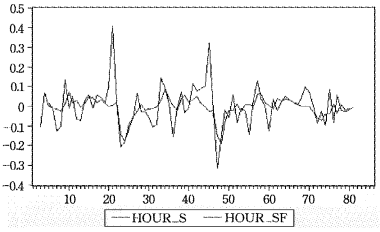

图 13　ARIMA(2,2,1)模型预测

从图 11 以及流量序列的长度特征可以看出采用小波变换与时间序列相结合的方式进行流量预测可以得到较好的预测结果，而图 13 则是使用 ARIMA 模型进行预测，得到的预测结果较次，但是该流量序列的平稳性以及自相似系数决定了使用 ARIMA 模型更加准确与方便，取得的预测结果是可以接受的。图 12 则是直接使用 FARIMA 模型拟合 BC-pAu89 数据集，未考虑到流量的自相似与序列的平稳性，得到的预测结果较差，不如使用小波变换与时间序列模型的效果。

本文所提算法的最大的优点就是根据自相似系数与平稳性两大特性选择小波变换与时间序列模型结合的方式，既保证了算法的灵活性，有效地减少了单一小波-FARIMA 模型的运算，又能够提高算法的预测准确度。

结束语　本文首先分析现有网络流量预测模型，并对网络流量的特征做了简要的分析，在此基础上提出一种基于流量特性与时间序列特性的流量预测算法。利用实际的网络流量进行模型拟合，通过对实验的结果分析，验证了所提模型的灵活性与预测结果的准确性。本文所提到的模型研究都是在有线网络流量上进行的，而无线网络与移动网络的网络结构与流量特点更加复杂，本文的研究方法将作为下一步研究无线网络的流量的基础。

参 考 文 献

[1] Leland W E, Taqqu M S, Willinger W, et al. On the self-similar nature of Ethernet traffic[J]. ACM SIGCOMM Computer Communication Review, ACM, 1993, 23(4): 183-193

[2] 高茜，冯琦，李广侠，等. 基于组合模型的自相似业务流量预测[J]. 计算机科学，2012，39(4)：123-126

[3] 马力，张高明，苟娟迎，等. 一种基于小波变换的校园网流量预测方法研究[J]. 计算机科学，2012，39(z2)：69-73

[4] Liu X, Fang X, Qin Z, et al. A Short-Term Forecasting Algorithm for Network Traffic Based on Chaos Theory and SVM [J]. Journal of Network and Systems Management, 2011, 19(4): 427-447

[5] Yu Y, Wang J, Song M, et al. Network Traffic Prediction and Result Analysis Based on Seasonal ARIMA and Correlation Coefficient[C]// Intelligent System Design and Engineering Application (ISDEA), 2010 International Conference on. IEEE, 2010: 980-983

[6] Wei X. Supporting vector-machine prediction of network traffic [C]// Electrical and Control Engineering (ICECE), 2011 International Conference on. IEEE, 2011: 3203-3206

(下转第 98 页)

从图中可以看出，当 k 逐渐变大时，匿名成功率越来越低，因为随着用户数量增多，组团匿名约束条件和最大值 k 约束使失败率升高。

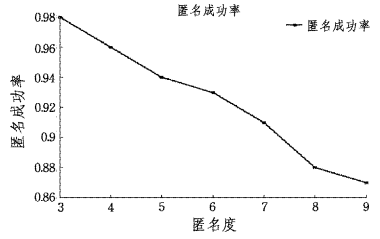

图 3　匿名度和匿名成功率关系

实验 2　两个模型的匿名处理时间

如图 4 所示，横坐标是匿名度 k，纵坐标是匿名时间。当 $k=3$ 和 4 时，两个模型匿名处理时间大致相同，因为此时都是第一次生成匿名区域；当 $5<k<9$ 时，两个模型的匿名处理时间上升，但是 CliqueCloak 模型用户等待时间总体较长，上升比较快。QR-TCM 模型采用了新的 CRCA 算法，虽然需要多次匿名，但每次匿名处理时间较短，因此缩短了用户等待时间，对于连续查询节省了大量时间。

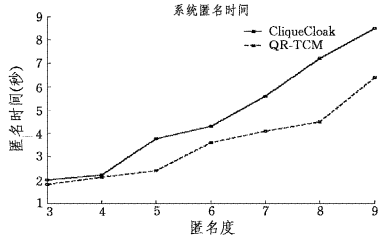

图 4　CliqueCloak 和 QR-TCM 匿名时间

实验 3　不同权重对 QR-TCM 模型的影响

表 3　A 组实验数据参数设置

参数名称	参数值	参数名称	参数值
P_1	—	w_1	0,5
P_2	0.7	w_1	0,1
P_3	—	w_1	0,2
P_4	0.6	w_1	0,1
P_5	0.5	w_1	0,1

表 4　B 组实验数据参数设置

参数名称	参数值	参数名称	参数值
P_1	—	w_1	0,1
P_2	0.5	w_2	0,2
P_3	—	w_3	0,5
P_4	0.5	w_4	0,1
P_5	0.8	w_5	0,1

如图 5 所示，横坐标是匿名度 k，纵坐标是两组不同参数下的位置服务质量。在表 3 和表 4 中，设置了不同的参数。从图中可以看出，随着匿名度增加，服务质量不断下降，A 组实验侧重隐私性，而 B 组实验侧重匿名处理时间，不同权重对 QR-TCM 模型影响比较大。

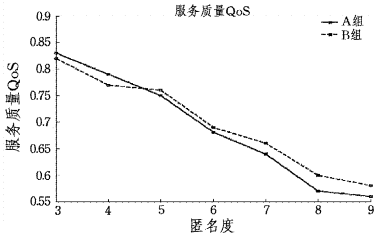

图 5　权重不同时，服务质量 QoS 不同

结束语　通过研究连续查询下的隐私保护问题，提出了 QR-TCM 用户隐私保护模型。该模型分析现有的位置隐私保护方法，发现位置服务延迟的缺点，提出了 CRCA 算法。该算法使得用户在使用位置服务信息时，能够在不降低匿名度的情况下，实时地得到精确位置服务。该模型平衡了匿名度和服务质量之间的矛盾，极大地提高了基于位置服务的质量。移动终端和服务器交互的过程中能量消耗比较多[11]，这些不足是我们下一阶段改进的重点。

参 考 文 献

[1] 潘晓，郝兴，孟小峰. 基于位置服务中的连续查询隐私保护研究[J]. 计算机研究与发展，2010，47(1)：121-129

[2] 王彩梅，郭亚军，郭艳华. 位置服务中用户轨迹的隐私度量[J]. 软件学报，2012，23(2)：352-360

[3] 王智慧，许俭，汪卫，等. 一种基于聚类的数据匿名方法[J]. 软件学报，2010，21(4)：680-693

[4] 魏志强，康密军，贾东，等. 普适计算隐私保护策略研究[J]. 计算机学报，2010，33(1)：128-138

[5] 周水庚，李丰，陶宇飞，等. 面向数据库应用的隐私保护研究综述[J]. 计算机学报，2009，32(5)：843-861

[6] 林欣，李善平，杨朝晖. LBS 中连续查询攻击算法及匿名性度量[J]. 软件学报，2009，20(4)：1058-1068

[7] 彭志宇，李善平. 移动环境下 LBS 位置隐私保护[J]. 电子与信息学报，2011，33(5)：1211-1216

[8] Pingley A，Wei Yu，Zhang Nan，et al. A context-aware scheme for privacy-preserving location-based services [J]. Computer Networks，2012，56(11)：2551-2568

[9] Lin Yu-bao，Chen Xiu-wei，Li Zhan，et al. An efficient method for privacy preserving location queries[J]. Front Computer Science，2012，6(4)：409-420

[10] 艾小淞，孙红，孙西国. SERVQUAL 和 SERVPERF 方法在 GPS 服务质量中的应用研究[J]. 北京航空航天大学学报：社会科学版，2010，23(4)：76-78

[11] Vergara-Laurens I J，Labrador M A. Preserving privacy while reducing power consumption and information loss in lbs and participatory sensing applications[C] // GLOBECOM Workshops (GC Wkshps)，2011 IEEE. 2011：1247-1252

(上接第 89 页)

[7] Zhao H，Ansari N. Wavelet Transform-based Network Traffic Prediction：A Fast On-line Approach[J]. Journal of Computing and Information Technology，2012，20(1)：15-25

[8] Maurya C K，Minz S. Fuzzy inference system for Internet traffic load forecasting [C] // Computing and Communication Systems (NCCCS)，2012 National Conference on. IEEE，2012：1-4

[9] 姜明，吴春明，张旻，等. 网络流量预测中的时间序列模型比较研究[J]. 电子学报，2009，37(11)：2353-2358

[10] http://ita. ee. lbl. gov/html/contrib

[11] http://datamarket. com/data/list/? q=time+series

[12] Mallat S G. A theory for multiresolution signal decomposition：the wavelet representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，1989，11(7)：674-693