文章编号：2095-2163(2019)01-0135-05 中图分类号：TP393.09 文献标志码：A

# 基于 ARIMA 和卡尔曼滤波的在线 Web 服务 QoS 预测方法

刘泽远，杨孝宗，舒燕君
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

**摘 要**：随着 Web 服务使用的广泛，人们普遍发现，Web 服务的服务质量（Quality-of-Service，QoS）受网络环境、服务端负载等诸多因素影响不断变化，而保证服务使用过程中的 QoS 也成为许多 Web 服务使用者的普遍要求。如何更好地帮助服务使用者选择未来一段时间内符合其服务质量要求的 Web 服务，同时也帮助服务提供者避免服务质量的违规，是服务计算领域近年来的热点问题。由于 ARIMA（Autoregressive Integrated Moving Average Model）模型参数简单并能较好地预测 QoS 违规，已经在 Web 服务的 QoS 预测领域获得了广泛的应用。但是单纯地使用 ARIMA 模型不能够适应 Web 服务 QoS 数据的波动频繁、包含噪声等复杂特点。为了达到更加准确的预测效果，本文提出了一种基于时间序列分析的 Web 服务 QoS 预测方法，该方法结合了 ARIMA 模型与卡尔曼滤波，对服务质量的波动反馈灵敏，较单一的预测模型能够有更准确的预测效果。

**关键词**：Web 服务；服务质量（QoS）；预测；ARIMA；卡尔曼滤波

## The Web services QoS prediction method based on ARIMA and Kalman filtering

LIU Zeyuan, YANG Xiaozong, SHU Yanjun
（School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China）

**【Abstract】** With the wide use of Web services, it is generally found that the QoS (quality of service) of Web services is constantly changing due to various factors such as the network environment and servers´ load. Business users commonly claim that QoS should be guaranteed. How to help service users select Web services that meet the quality of service requirements for a period of time in the future, and also help service providers to avoid service quality violations are the hot issue in the field of service computing in recent years. This paper proposes a Web service QoS prediction method based on time series analysis. This method combines ARIMA model and Kalman Filtering. Meanwhile it is sensitive to the fluctuation of service quality, and can have more accurate prediction effect than a single prediction model.

**【Key words】** Web Services; QoS; prediction; ARIMA; Kalman Filtering

## 0 引 言

Web 服务为面向服务的体系架构（Service-Oriented Architecture，SOA）提供了一个标准化的解决方案[1]，并在设计上成功解决了分布式计算基于标准、松耦合、协议无关的需求。随着 Web 服务的指数级增长，研究发现很多服务的服务质量良莠不齐，仅仅通过功能上的要求无法做出有效选择，因而需要通过服务质量（Quality of Service，QoS）的评估，来选择满足要求的 Web 服务。服务质量描述 Web 服务的非功能方面的特性，主要包括响应时间、吞吐量、故障间隔时间等指标。

在波动大、变化快的分布式 Web 服务环境中，Web 服务的服务质量不会保持不变，而要保证 Web 服务满足使用者的要求，就要寻找一种有效的方法来判断 Web 服务在一定时间内能否满足服务质量的约束。针对这一问题，研究指出可以对 QoS 历史记录数据进行分析，用来预测 Web 服务在未来时间的 QoS 值。立足于这一研究角度，很多有关基于时间序列分析的 Web 服务 QoS 预测方法的研究已陆续涌现[2]，同时也证明了这一思路的设计可行性。

本文在已有研究基础上提出一种基于时间序列分析的 Web 服务 QoS 预测方法，结合了 ARIMA（Auto-regressive Integrated Moving Average）与卡尔曼滤波（Kalman Filtering）技术，首先使用 ARIMA 模型对 QoS 历史记录值进行拟合，将 ARIMA 模型转换成状态空间模型，使用卡尔曼滤波器预测未来时间点的 QoS 值，在运行时每次测得的 QoS 值都可以对卡尔曼滤波器进行更新。该方法具有适应性高，计算复杂度低等特点，通过在公开的 Web 服务 QoS 记录数据集[2]上实施的实验可以验证该方法的准确性。

作者简介：刘泽远（1995-），男，硕士研究生，主要研究方向：服务计算；杨孝宗（1939-），男，博士，教授，博士生导师，主要研究方向：计算机系统结构、容错计算技术、故障注入技术等；舒燕君（1981-），女，博士，讲师，硕士生导师，主要研究方向：计算机系统结构、软件可靠性、服务计算等。

收稿日期：2018-06-11

本文研究在论述上,首先探讨了该方向的相关研究工作,并分析了现有工作的优缺点;接下来给出了本文预测方法的理论基础与具体设计;然后,将该方法与经典的方法展开比较实验,并对实验结果进行分析;最后总结全文,进而展望未来的研究工作。

## 1 相关研究

在基于时间序列分析的 Web 服务的 QoS 预测中,目前已可见到一系列的研究工作和学术成果。Cavallo 等人[2]比较了在时间序列模型中堪称代表的 ARIMA 模型与当前值方法、历史平均值方法以及线性回归法 AR 的仿真测试,结果表明 ARIMA 算法对异常值的容忍度较高,并且对 QoS 的违规有良好的预测效果,在预测结果上能够表现出相对较小的预测误差。Amin 等人[3]提出结合 ARIMA 与广义自回归条件异方差(GARCH)模型,GARCH 模型可以改善 ARIMA 模型中方差恒定假设的不足,更加贴近真实的 Web 服务 QoS 数据特征。Ye 等人[4]考虑 QoS 多个属性维度之间的相关性,提出使用多元 ARIMA 模型、Holters-Winters 指数平滑模型,并在此后的研究中,将提出的方法同 VAR(向量自回归)做了比较处理,其中包括了许多基于自回归的方法,诸如 AR、SETAR(自激励门限回归模型)、ARIMA 以及 ARMA-GARCH 模型。

可以发现,将单一的统计学模型进行组合是研究 Web 服务 QoS 预测的热门趋势,单一的模型已难以满足波动大并且没有统一规律的 Web 服务 QoS 预测。而随着近年来机器学习技术的迅猛发展,已经有一些研究者开始尝试将机器学习的方法应用到 Web 服务 QoS 的预测中。Zadeh 等人[5]和 Senivongse 等人[6]即应用 ANN(人工神经网络)解决 Web 服务 QoS 预测的问题,Yang 等人[7]则提出使用遗传编程的方法。Wang 等人[8]提出使用长短期记忆循环神经网络模型,预测 Web 服务系统的可靠性。综合分析上述研究结果可知,机器学习的方法并未能在预测准确率上达到质的提升,而这些方法在执行上的复杂却已经成为将其付诸现实应用的制约因素。

在使用 ARIMA 模型预测 Web 服务的 QoS 时间序列的研究中经过分析发现,单纯地使用 ARIMA 模型不能够适应 Web 服务 QoS 数据的波动频繁、包含噪声等复杂特点,为了获得更加准确的预测效果,本文使用卡尔曼滤波对 ARIMA 模型的结果进行修正。卡尔曼滤波是一种最优化自回归数据处理算法(Optimal Recursive Data Processing Algorithm)[9],其最突出的优势即在于能够从一些包含噪声的观测量估计系统的状态。卡尔曼滤波利用前一时刻的估计值和当前时刻的观测值,更新当前时刻的状态向量,利用这一特点就可以很好地弥补单一 ARIMA 模型在 QoS 时间序列预测上的不足。下面首先详述 ARIMA 与卡尔曼滤波相结合的理论基础,然后阐述该方法的设计过程。

## 2 理论基础

ARIMA 模型与卡尔曼滤波相结合的关键就是将 ARIMA 模型转换成状态空间模型,再使用卡尔曼滤波对状态进行预测与更新。一个 ARIMA( $p$,$d$, $q$ )模型经 $d$ 阶差分就可得到 ARMA ($p$,$q$) 模型,数学公式可表述如下:

$$Z_t = \sum_{i=1}^{p} \phi_i Z_{t-i} + a_t - \sum_{j=1}^{q} \theta_j a_{t-j} \tag{1}$$

其中, $Z_t$ 为 $t$ 时刻观测值; $a_t$ 为 $t$ 时刻预测值与观测值的误差; $\phi_i$、$\theta_j$ 为模型参数。设 $m = \max\{p, q\}$,令$\phi_i = 0(i > p)$,$\theta_j = 0(j > q)$,那么公式(1)也可以表示为:

$$\boldsymbol{Z}_t = \sum_{i=1}^{m} \phi_i \boldsymbol{Z}_{t-i} + \boldsymbol{a}_t - \sum_{j=1}^{m} \theta_j \boldsymbol{a}_{t-j} \tag{2}$$

设 $B$ 为延迟算子,可将公式(2)变化为:

$$\phi(\boldsymbol{B})\boldsymbol{Z}_t = \theta(\boldsymbol{B})\boldsymbol{a}_t \tag{3}$$

接下来,要将式(3)转换为 $\boldsymbol{Z}_t = \psi(\boldsymbol{B})\ \boldsymbol{a}_t$ 的形式,令 $\psi(\boldsymbol{B}) = \theta(\boldsymbol{B})/\phi(\boldsymbol{B})$,其中 $\psi(\boldsymbol{B}) = (\psi_0 + \psi_1 \boldsymbol{B} + \cdots + \psi_m \boldsymbol{B}^m + \cdots)$,$\psi_0 = 1$。根据延迟算子 $\boldsymbol{B}$ 的每一项系数相等,可以得出:

$$-\theta_m = -\phi_m \Psi_0 - \phi_{m-1} \Psi_1 - \cdots - \phi_1 \Psi_{m-1} + \Psi_m \tag{4}$$

化简公式(4),就会得到 $\psi_m$ 的数学运算公式可见如下:

$$\Psi_m = \sum_{i=1}^{m} \phi_i \Psi_{m-i} - \theta_m \tag{5}$$

由此,可以将 $\boldsymbol{Z}_{t+m-i} = \psi(\boldsymbol{B})\ \boldsymbol{a}_{t+m-i}$ 表达式展开为如下形式:

$$\boldsymbol{Z}_{t+m-i} = \boldsymbol{a}_{t+m-i} + \Psi_1 \boldsymbol{a}_{t+m-1} + \Psi_2 \boldsymbol{a}_{t+m-i-2} + \cdots \tag{6}$$

$$\boldsymbol{Z}_{t+m-i|t} = \Psi_{m-i} \boldsymbol{a}_t + \Psi_{m-i+1} \boldsymbol{a}_{t-1} + \Psi_{m-i+2} \boldsymbol{a}_{t-2} + \cdots \tag{7}$$

$$\boldsymbol{Z}_{t+m-i|t-1} = \Psi_{m-i+1} \boldsymbol{a}_{t-1} + \Psi_{m-i+2} \boldsymbol{a}_{t-2} + \cdots \tag{8}$$

由公式(7)、(8)可得:

$$\boldsymbol{Z}_{t+m-i|t} = \boldsymbol{Z}_{t+m-i|t-1} + \Psi_{m-i} \boldsymbol{a}_t \tag{9}$$

令状态向量 $\boldsymbol{S}_t = (Z_{t|t-1}, Z_{t+1|t-1}, \ldots, Z_{t+m-1|t-1})^{\mathrm{T}}$,

观测值 $\boldsymbol{Z}_t = \boldsymbol{Z}_{t|t-1} + \boldsymbol{a}_t$，观测方程即可表示为：

$$\boldsymbol{Z}_t = [1,0,\cdots,0]\boldsymbol{S}_t + a_t \tag{10}$$

根据公式(5)、(9)可推得状态转移方程为：

$$\boldsymbol{S}_{t+1} = \boldsymbol{F}\boldsymbol{S}_t + \boldsymbol{G}a_t \tag{11}$$

其中，$\boldsymbol{F}$、$\boldsymbol{G}$ 的运算可使用如下数学公式：

$$\boldsymbol{F} = \left[\begin{array}{c|cccc} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & & \\ \phi_m & \phi_{m-1} & \cdots & \phi_2 & \phi_1 \end{array}\right] \quad \boldsymbol{G} = \begin{bmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \\ \vdots \\ \Psi_m \end{bmatrix} \tag{12}$$

综上就是模型 ARMA$(p,q)$ 转移到状态空间模型的推导过程，在得到状态空间模型后，就可转入卡尔曼滤波的应用设计的研究。

## 3　方法设计

如图 1 所示，本文提出的 Web 服务 QoS 预测方法主要分为 2 个模块，分别是：ARIMA 模型的确立和卡尔曼滤波器的更新。总地来说，根据 QoS 历史数据进行 ARIMA 建模，根据第 2 节的方法将其转换为状态空间模型，在运行时获得新的 QoS 观测值后，对状态向量进行更新，再将卡尔曼滤波的预测值作为最后的预测结果。对此内容，本节将给出阐释分述如下。

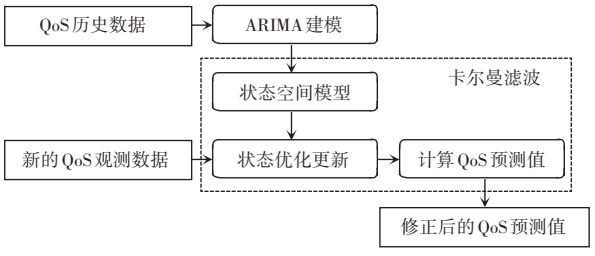


**图 1　预测方法概览**

**Fig. 1　Overview of the proposed prediction method**

### 3.1　ARIMA 模型建立

(1)白噪声检测。如果一个时间序列是纯随机的，此时就没有必要再利用 ARIMA 方法继续接下来的操作，因其并不适用于该方法。若一个时间序列纯由白噪声构成，那么在理论上其各阶自相关系数(Auto-Correlation Function，ACF)均为 0，基于这一点即可判断时间序列是否为纯随机。

(2)稳定性检测。最简单的判断时间序列稳定性的方法就是通过肉眼观察样本的二维曲线图得出结论；当然也有其它的方法。理论上可证明平稳的时间序列通常会具有短期相关性，随着延迟增加，序列的自相关系数很快会降低至 0，根据这一点就能判断时间序列的平稳性。如果序列非平稳，则引入逐次差分，直至得到平稳序列，即将 ARIMA 模型转化成 ARMA 模型。

(3)定阶。得到平稳的时间序列后，需进一步判断序列是否满足自回归(AR)、滑动平均(MA)。识别方法是序列的自相关系数(ACF)和偏自相关系数(Partial Auto-Correlation Function，PACF)。若 ACF 曲线衰减的同时、PACF 曲线截断，则 AR 模型适用；若 ACF 曲线截断的同时、PACF 曲线衰减，MA 模型适用。根据 ACF 的拖尾特征、PACF 的截尾特征，可以确定对应的阶数 $p$、$q$。

(4)模型参数拟合与检验。模型的阶数确定后，模型参数个数也确定了。需要拟合出其它参数，即 $\phi_i(i=1,2,\ldots,p)$、$\theta_j(j=1,2,\ldots,q)$，文献[10]给出使用极大似然估计法进行参数拟合的完整计算步骤。对于模型的检验环节，就需要检验模型是否具有统计意义，即检验是否对时间序列提取足够充分的信息。理论上推导可知，如果模型信息提取充分，残差序列即为白噪声。

### 3.2　卡尔曼滤波预测

卡尔曼滤波分为时间更新方程和状态更新方程。在 Web 服务的 QoS 预测中，结合 ARMA$(p,q)$ 转化的状态空间方程，设 $\boldsymbol{S}_{t|t}$ 为时刻 $t$ 基于先验信息得到的状态估计，卡尔曼滤波时间更新方程的公式表达详见如下：

$$\boldsymbol{S}_{t+1|t} = \boldsymbol{F}\boldsymbol{S}_{t|t} \tag{13}$$

$$\boldsymbol{P}_{t+1|t} = \boldsymbol{F}\boldsymbol{P}_{t|t}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{G}\boldsymbol{Q}\boldsymbol{G}^{\mathrm{T}} \tag{14}$$

其中，$\boldsymbol{P}_{t|t}$ 是时刻 $t$ 预测误差的后验协方差矩阵；$\boldsymbol{P}_{t+1|t}$ 是时刻 $t+1$ 预测误差的先验协方差矩阵；$\boldsymbol{Q}$ 为过程噪声的协方差矩阵。状态转移矩阵 $\boldsymbol{F}$ 将 $\boldsymbol{S}_{t|t}$ 转换为 $\boldsymbol{S}_{t+1|t}$，将 $\boldsymbol{P}_{t|t}$ 转换为 $\boldsymbol{P}_{t+1|t}$。$t+1$ 时刻 QoS 预测值根据状态向量 $\boldsymbol{S}_{t+1|t}$ 可由如下公式计算得出：

$$\boldsymbol{Z}_{t+1|t} = \boldsymbol{H}\boldsymbol{S}_{t+1|t} \tag{15}$$

得到观测值 $\boldsymbol{Z}_{t+1}$，可以对 $\boldsymbol{S}_{t+1|t+1}$、$\boldsymbol{P}_{t+1|t+1}$ 做出更新，卡尔曼滤波的状态更新方程如式(16)所示：

$$\boldsymbol{S}_{t+1|t+1} = \boldsymbol{S}_{t+1|t} + \boldsymbol{P}_{t+1|t}\boldsymbol{H}^{\mathrm{T}}[\boldsymbol{H}\boldsymbol{P}_{t+1|t}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R}]^{-1}(\boldsymbol{Z}_{t+1} - \boldsymbol{Z}_{t+1|t}) \tag{16}$$

$$\boldsymbol{P}_{t+1|t+1} = \boldsymbol{P}_{t+1|t} - \boldsymbol{P}_{t+1|t}\boldsymbol{H}^{\mathrm{T}}[\boldsymbol{H}\boldsymbol{P}_{t+1|t}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R}]^{-1}\boldsymbol{H}\boldsymbol{P}_{t+1|t} \tag{17}$$

其中，$\boldsymbol{P}_{t+1|t}\boldsymbol{H}^{\mathrm{T}}[\boldsymbol{H}\boldsymbol{P}_{t+1|t}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R}]^{-1}$ 就是卡尔曼增益矩阵，通过最小化预测误差的后验协方差矩阵计算得出，计算结果决定了状态向量受到观测值的影响程度。

在实际操作中，输入初始状态 $S_{0|0}$ 和 $P_{0|0}$，然后计算 $Z_{1|0}$ 和 $V_{1|0}$。当得到观测值 $Z_1$ 后，使用更新方程计算 $S_{1|1}$ 和 $P_{1|1}$，将其作为下一次循环的初始状态。值得一提的是，输入任意的初始值 $S_{0|0}$ 和 $P_{0|0}$，初始值对后面预测值的影响会随着 $t$ 的增加而变得越来越小，因为状态转移矩阵 $\boldsymbol{F}$ 的特征值均小于 1，也就是说，随着 $t$ 的增加，卡尔曼滤波器保证了初始值对后面结果的影响将逐步趋近于 0。

## 4 实验

本节将验证研发提出的预测方法的实际效果。实验数据采用 Cavallo 公开发布的数据集[2]，选取了 XML Daily Fact 的服务，该数据记录了 2006 年 7 月～11 月对该服务每隔 1 h 调用一次的 QoS 值记录，包括响应时间、可用性等。本次测试实验拟对实践中最常用的响应时间属性进行分析，选取了连续 400 个时刻的记录作为实验数据。为了削减异常值对模型的影响，对实验数据中高于 3 000 ms 的值（这些值相比样本个数非常少）设为正常数据的均值。

首先是使用单一的 ARIMA 模型的预测效果，绘制后即如图 2 所示。

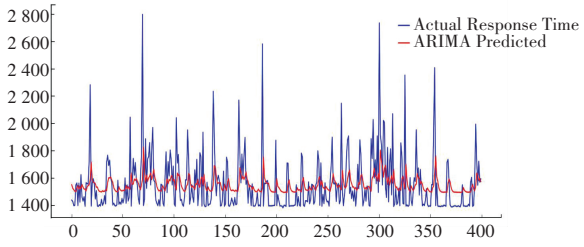


**图 2 ARIMA 预测值与实际值**

**Fig. 2 ARIMA predicted value vs. actual value**

使用本文提出的预测方法，结合 ARIMA 模型与卡尔曼滤波的预测效果则如图 3 所示。

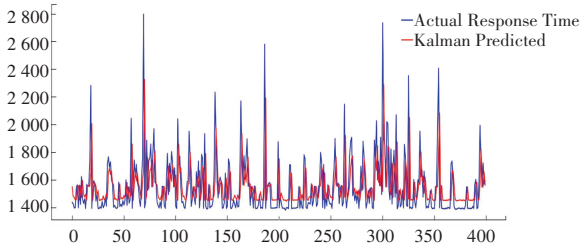


**图 3 卡尔曼滤波预测值与实际值**

**Fig. 3 Kalman filtering predicted value vs. actual value**

从图 2、图 3 中可以直观地看出，本文的预测方法的预测结果更加接近真实值，对波动情形的反应更为及时，在下一时刻就能做出调整与修正。为了更加客观地展现本文预测方法的准确率，对 2 种预测方法的预测结果与实际值的误差进行统计。本文计算了预测结果与实际值的均方根误差（Root Mean Squared Error，RMSE），均方根误差的数学定义可表示如下：

$$RMSE = \sqrt{\frac{\sum_i (r_i - \hat{r}_i)^2}{N}} \tag{18}$$

其中，$r_i$ 为时刻 $i$ 的实际值；$\hat{r}_i$ 为时刻 $i$ 的预测值；$N$ 为样本序列长度。

基于单一的 ARIMA 模型预测结果，均方根误差为216.862 6；基于本文预测方法，预测结果的均方根误差为200.673 3，预测效果整体提升了 7.47%。

## 5 结束语

准确地预测 Web 服务的 QoS，能够为服务提供者有效地降低服务质量违规的风险，而且也能够帮助服务使用者在使用时间内调取到服务质量稳定的服务。故而，本文研发设计了一种基于时间序列分析的 Web 服务预测方法，并在公开的数据集上进行实验，验证了该方法的有效性。实验结果表明，相比传统的单一 ARIMA 模型预测方法，本文方法能够自适应地实现对 QoS 波动的预测，进而及时发出 QoS 违规的预警。

Web 服务的 QoS 相对其它领域的时间序列有其本身的特点，由于业务复杂程度不一、网络环境波动大的影响，要更加准确地预测 Web 服务的 QoS 值就势必还有很多的工作需要成为当下的关键研究课题。后续工作将主要着重于如下方面：

（1）时间序列的噪声。噪声会对序列本身信息的挖掘产生影响，而去噪方法的选择也将涉及多方面的因素考证，因此为 Web 服务的 QoS 序列进行合理去噪，将会是未来的研究热点。

（2）QoS 各属性的相关性。单变量的预测方法更加难以抵抗噪声对预测结果的影响，如何挖掘多个属性之间的相关性，提高预测准确率，则亟待后续的深入系统研究。

## 参考文献

[1] PAPAZOGLOU M P, VAN DEN HEUVEL W J. Service oriented architectures: Approaches, technologies and research issues[J]. The VLDB Journal, 2007, 16(3): 389-415.

[2] CAVALLO B, PENTA M D, CANFORA G. An empirical comparison of methods to support QoS-aware service selection[C]//Proceedings of the 2nd International Workshop on Principles of Engineering Service-Oriented Systems. Cape Town: ACM, 2010: 64-70.

（下转第 142 页）