

基于自回归技术的自适应服务质量测量方法^{*}

赵成栋

(北京航空航天大学软件开发环境国家重点实验室 北京100083)

摘 要 SLA 管理已经成为服务商和用户关心的重要内容。在进行服务质量管理时,服务质量的测量是其他管理功能的基础。本文从采样的角度分析了服务质量测量和传统网络测量的异同点,给出了服务质量采样频率的理论下限。在讨论已有方法不足的基础上,提出了基于预测的自适应采样算法,该算法在采样区间内随机确定采样的具体时间点,并结合自回归预测方法对即将采样的结果进行预测,在此基础上以一定的概率接受预测结果,放弃对准确预测时间点的采样,从而减少了采样次数,降低了对服务的影响。最后的实验结果说明了算法的有效性。

关键词 SLA 服务质量,采样,自回归,预测,服务测量,网络测量

An Adaptive Method for Service Level Monitoring Using Auto-Regressive Technique

ZHAO Cheng-Dong

(National Laboratory of Software Development Environment, BeiHang University, Beijing 100083)

Abstract SLA (Service Level Agreement) and its management have become more and more important for both service providers and customers, within which measurement of service level is fundamental for other managements. This paper discusses difference between service measurement and traditional network measurement from sampling perspective, and deduces a theoretical lower-limit on sampling frequency. To reduce the sampling and keep high accuracy of service level data, the paper presents an adaptive algorithm based on estimation method, which determines time to sample using random function during a small even interval, then constructs an auto-regressive model according to historical data to estimate the current data of service at the sampled time, accepts the result by an adaptive probability method, and gives in sampling at this interval. The algorithm decreases sampling's influence on measured service through reducing sampling times. Experiments show its good efficiency.

Keywords Service level agreement, Sampling, Auto-regressive, Estimation, Service measurement, Network measurement

1. 引言

随着网络技术的高速发展和人们对网络服务的需求不断增加,网络及其相关的应用越来越多地以服务的形式提供给大众。这种服务商-用户之间的关系正在逐步转向依赖服务水平协议(Service Level Agreement)的管理来维系。所谓 SLA 是在一定开销下(通常这个开销是驱动提供有保障的服务质量的主要因素),为了保障服务的性能和可靠性,服务提供商和用户间或者服务提供商之间定义的一种双方认可的协定。它规定了服务等级和服务所必须满足的性能水平(包括服务水平测量、服务水平报告和信誉及费用三个方面),并使服务提供商有责任完成这些预定的服务等级^[1]。服务商必须按照 SLA 中指定的质量水平给用户提供服务。

在服务水平的管理中,首要的是对服务水平的测量。服务水平的检测方法和传统的网络测量^[2]的方法类似,可以分为主动测量(Active Measurement)和被动测量(Passive Measurement)。主动测量是通过向网络、服务通道和目标注入一定的数据,测定由此引发的流量、应答、行为等特征,从而获得服务的质量特性;被动测量是使用硬件、软件的探针(Probe)或者代理(Agent)捕捉、检索、跟踪网络、服务通道和目标的行为数据,进行进一步分析来获得服务质量的情况。

考虑到服务的普遍复杂性,一般很难描述,也很难从网络

报文中定位和辨别一个服务的过程和行为特征,因此被动测量很难完成服务质量水平的测量;而采样主动测量的方法,可以通过实现测量代理的方法,精确区分服务建立、请求、应答等所有活动的每个细节,从而可以正确地评估服务的质量水平。

值得指出的是,服务水平的测量和网络测量有一些较大的区别。网络测量是对一个网络及其较大范围内的指标进行测量,即使采用主动测量的方式,其发生的流量和对原有网络的行为的影响都很小;而服务的测量则不同,用户购买服务是以完成一定的任务为目标的,相应地,出于对成本和利用率的考虑,购买的服务一般只要保证满足最大需求的一个上限。可以设想,一个服务商要提供这样大量的低要求(低带宽、低响应次数、服务有限)的服务,一次主动测量可能对这个服务商的整个网络影响微乎其微,但是对于单个的用户来讲,影响就不可忽视,甚至耗尽服务的总容量。所以,对服务进行主动测量要考虑两个因素:

(1)主动测量是否相对整个服务的影响限制在一个可以接受的范围内;

(2)主动测量不但要尽量能够真实的反应实际服务的质量水平,同时要最大限度地降低对服务的影响。

本文的以下部分针对以上两个因素讨论了服务质量水平测量的方法,提出了一个自适应采样算法,提高了效率。第2节

^{*} 本课题得到国家九七三项目(“网络环境下海量信息组织与处理的理论与方法研究”中的“基于先进网络的大型电信网络管理示范系统”子项目)的资助(G1999032709)。

对传统采样方法进行了讨论,并给出了采样次数的下限;第3节给出了自适应算法的算法描述,并通过实验证明了算法的有效性。

2. 服务质量采样方法与采样频率

2.1 传统采样方法

主动测量每次实施测量的结果,都可以看作对该服务质量水平的一次采样。通过对多个样本的统计和分析,就可以得出服务质量水平的近似结果。显然,采样的频率和时机都会对采样的结果及最终服务质量水平的评估有决定性的影响。常见的采样方法有以下几种^[3-6]:

系统采样 是使用一个确定的函数选择开始采样的时间点和采样的时间长度进行采样。

随机采样 是在采样区间里根据随机函数的结果选择采样(起始)点。这样,每次采样都是一次独立实验,得到的也就是一个无偏的估计。和系统采样的不同在于,要使用随机函数来获得随机数。具体的采样方法包括:

(1)多点采样(n-out-of-N sampling):在多点采样中,是要从一个包含N个元素的总体中,选择n个元素作为样本。在多点采样中,采样的次数,即获得的样本的容量是固定的。

(2)概率采样(Probabilistic sampling):概率采样要依据一个预先定义的选择概率来确定一个元素是否被采样。这种采样方法的采样次数是不确定的。

分层采样 分层采样是把采样过程分成多个步骤。首先,总体中的元素根据某个给定的特性分成多个子集。这个分组过程也可以多步完成。然后从每个子集中进行采样。

2.2 服务质量水平与采样

一般地,在SLA的协议中,将规定服务的质量参数及其预期水平。具体对每种不同的服务、每个不同的顾客要求都可能差别很大,但是以下3个参数对大多数情况都是有效的:

服务可得性:在指定的时间区间内,或者某个时刻,服务是否可以获取。

服务响应时间(SRT):服务响应请求或者完成请求所经历的时间。

服务吞吐量:服务实际提供的次数、流量、速率等特征。

在服务水平的表述和验证上,也有和以往网络测量不同的地方。一个典型的服务水平协议可能规定如下的服务质量水平:

服务质量水平参数协定

购买专线接入服务,带宽为10M,要求:

- 接通率不低于99%,
- 双向网络延时在90%的时间内低于20ms,
- 丢包率在流量为8mbps以下时90%的时间内小于5%
-

我们把“接通率为不低于99%”换一种表述方法:服务在指定时间段内以99%的概率可得。假如服务时间为一个月,那么服务中断的时间总长度不得超过:

$$\begin{aligned} & \text{一个月时间} \times (1 - 99\%) = 24 \text{小时/天} \times 30 \text{天} \times 1\% \\ & = 7.2 \text{小时} \end{aligned}$$

也就是说,所有服务中断的时间的总和不得大于7.2小时。可以看到,即使对服务质量水平的某一个参数指标的统计也不再是简单的频率累加计算,而是和时间结合在一起的二维的分析处理,一个样本点代表一个服务水平状态的开始或者结束,而不是简单的服务水平的特征表现。

2.3 服务质量的采样频率

如前所述,采样的频率也是影响采样有效度的主要因素之一。下面定量地讨论样本容量的限制^[7]。

设样本容量为 n , $X_i (i=1, 2, \dots, n)$ 是 n 个样本。对于质量指标 V ,可以通过样本的数学期望 $\hat{V}$ 来考察:

$$\hat{V} = \frac{1}{n} \sum_{i=1}^n X_i$$

$\hat{V}$ 是 V 的无偏一致估计。为了保证样本能够有效地反应总体的情况,我们可以限制样本相对误差在一个可以接受的范围内,例如:给定参数 $(\epsilon, \eta) (0 < \epsilon, \eta < 1)$,令

$$P\left\{\left|\frac{\hat{V}-V}{V}\right| > \epsilon\right\} \leq \eta \quad (1)$$

即我们希望样本相对误差小于 ϵ 的概率大于 $1-\eta$ 。

假定总体均值为 μ ,标准方差 σ 。根据中心极限定理,当样本容量 $n \rightarrow \infty$, (1)式可以进行如下变换:

$$\begin{aligned} P\left\{\left|\frac{\hat{V}-V}{V}\right| > \epsilon\right\} &= P\left\{\left|\frac{\hat{V}-V}{\sigma}\right| > \frac{\epsilon\mu\sqrt{n}}{\sigma}\right\} \\ &\approx 2(1 - \Phi\left(\frac{\epsilon\mu\sqrt{n}}{\sigma}\right)) \leq \eta \end{aligned}$$

$$\text{可以推出: } n \geq \left(\frac{\Phi^{-1}(1-\eta/2)\sigma}{\epsilon\mu}\right)^2 \quad (2)$$

其中, μ 和 σ 可以使用大量以往测量获得的大量样本的无偏估计值进行近似计算。

由(2)式可以得知,当我们考虑一个质量指标对采样频率的影响时,通过指定样本相对误差 ϵ 在置信水平 η 内变动,那么可以确定最少样本容量的下限;另一方面,采样次数当然越大越好,理论上进行连续的无穷采样,但是实际上显然是不可行的,具体的取值要结合服务质量水平的要求精度、服务本身能够达到的最大上限以及采样过程的可行性来确定。

2.4 已有方法的不足

根据以上对服务质量水平测量特性的讨论,我们知道:对于服务质量水平的采样测量,不仅需要足够的采样次数获得足够的样本来考察总体的情况,而且还要采样次数尽可能地少,同时又要能够反映每个较小时间区间上的质量分布情况。但是,已有的采样方法不能很好地满足服务测量的需要:

• 系统采样通过周期性的采样虽然可以均等地了解每个时间区间上的服务水平,但是如果服务质量水平在某个特定的时间段内也具有一定的周期性表现,系统采样将无法获得这一情况。例如服务每5分钟左右出现一次服务短时间的中断又立刻恢复,而系统采样也是5分钟一次采样,由于都是确定的同一个偏移点的采样,发现的可能性几乎没有。事实上,在实际系统中,由于系统、线路设备、信号收到噪声等因素的影响,这种情况在一定程度上是普遍存在的^[8]。

• 随机采样根据随机数选择采样的时间点,更加具有不确定性,虽然可以更容易检测到周期性质量水平的变动,但是两次采样的时间间隔很大时无法反应该时间区间上的质量情况(under-sampling),很小时又是不必要的密集采样(over-sampling),因此也是不理想的采样方法。

• 分层采样在实际样本未产生之前难以进行分组,在主动采样的环境下也不够有效。

本文提出了一种基于自回归技术的自适应采样算法,把以上采样方法的优点结合在一起,又避免了它们的缺陷,可以很好地满足服务质量水平采样的要求。

3. 自适应采样算法

3.1 算法概述

首先,根据对一个特定服务S的质量水平要求,将整个时间区间均匀分成多个小的区间,每个区间内最多采样一次。区

间的划分满足:

(I) 区间的个数即采样的个数是满足(2)式的一个整数值;

(II) 区间的大小满足质量水平要求的精度。比如对可得性进行采样,中断的时间长度为 T_{down} , 区间长度 $T_{interval}$ 应该远小于 T_{down} 。

其次,在每个区间内选择采样的点。由于随机采样算法可以更好地反应该区间的质量水平,因此在每个区间内使用随机算法,确定要采样的时间点。刚才已经讨论过,服务采样的次数要尽可能地少。对采样区间个数的划分由于考虑到多种因素,显然不能取定计算得到的下限值。在实际进行采样时,还需要有效的方法来减少采样的次数。

本文的想法是,通过一种自适应的基于预测的技术,使用前面测量获得的数据预测当前要采样区间采样点的值的情况,以某种概率接受这个预测,如果预测被接受,我们就认为当前区间采样点的情况和预测结果基本一致,放弃对该区间进行采样。如果不接受这个预测,就在采样点进行采样,同时,根据预测的结果和实际采样结果的情况自适应地调整采样策略参数。转到下一个区间,重复这个过程。时间序列的预测方法比较多,如中值法、滑动平均法、自回归法等,通过对历史采样数据的模型构造和拟合,选择了自回归方法进行预测。

3.2 自回归模型

自回归模型(Auto-Regressive Model)是进行时间序列分析常用的一种方法,通过对以往采样数据的多项式计算预测当前的数据。

定义(p 阶自回归过程) 时间序列 $\{Y_t\}$ 称为 p 阶自回归过,如果满足

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \epsilon_t$$

其中 c 为常数, ϕ_i 为常数, $\{\epsilon_t\}$ 为白噪声序列, $p \in N$ 。

该模型的主要优点是:(1)AR 模型直观,易于理解,容易掌握,计算效率高;(2)各种因素对预测目标的印象是通过它们在时间过程中的综合体现被考察的,并将序列的历史观察值,作为诸多因素影响与作用的结果。

我们把对同一个服务不同时刻的测量结果可以看作一个时间序列,在历史数据的基础上进行 AR 建模,确定模型的阶数(即 p 的大小),进而通过 Yule-Walker 等方法求出 ϕ_i 。这样,通过对 Y 的迭代计算,就可以不断地预测该序列在当前时间的取值,为优化采样提供依据。

3.3 算法描述

现在假设对服务 S 进行服务水平的测量,测量参数为 Y , SLA 协议规定了 Y 满足条件 $C(Y)$ (如 $Y < Y_{threshold}$) 为期望达到的质量水平。本文提出的采样算法如下:

(1)根据历史采集的数据建立 $AR(p)$ 预测模型:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} \quad (\text{忽略白噪声}) \quad (3)$$

(2)根据质量水平精度确定采样周期 $T_{interval}$ 的大小。

(3)进入初始采样期:在采样区间到来时,采用随机函数确定采样的时刻,在采样时刻进行采样,获得采样值 y_k^{real} ,同时,根据(3)式计算预测值 $y_k^{estimation}$ 。比较 y_k^{real} 和 $y_k^{estimation}$,记录预测的准确性。使用下式来描述预测的准确性:

$$e_k = \begin{cases} 1, & \text{如果 } \frac{|y_k^{estimation} - y_k^{real}|}{y_k^{real}} < \alpha \text{ 成立} \\ 0, & \text{如果 } \frac{|y_k^{estimation} - y_k^{real}|}{y_k^{real}} \geq \alpha \text{ 成立} \end{cases}$$

即判断预测的相对误差如果在一个较小的范围内,那么

我们就认为是一次准确的预测。其中 α 是预先设定的可以接受的误差变动范围,如 $\alpha = 0.1$ 。

如此采样 m 次, m 是一个较大的整数。以后进入预测采样期。

(4)进入预测采样期:当第 i 个采样周期到来时

(4.1)使用(3)式进行预测当前周期内的采样值 $y_i^{estimation}$ 。

(4.2)计算预测准确度函数 $V(i)$: $V(i) = (\sum_{j=1}^{i-1} e_j) / (i-1)$

(4.3)如果 $y_i^{estimation}$ 满足质量条件 $C(Y)$,以概率 $p = V(i)$ 接受该预测结果,即产生一随机数 r ,如果 $p < r$,本采样区间不进行采样,转(4.1);如果 $p \geq r$,转(4.5)。

(4.4)如果 $y_i^{estimation}$ 不满足质量条件 $C(Y)$,以概率 $p = 0$ 接受该预测结果,即不接受该预测结果,转(4.5)。

(4.5)产生一个随机数,确定本区间内采样的时刻点。在该时刻到来时进行采样。

(4.6)记录本次预测结果和实际采样结果,计算 e_i 。转(4.1)。

3.4 实验结果

为了获取服务的真实质量情况并方便对各种不同采样方法的比较分析,本文对一个 HTTP 服务 S_{http} (<http://e.pku.edu.cn/>)进行了连续14天的密集(2分钟一次)采样,获得8000个左右的数据。以实际测量的服务响应时间(记从发送对一个页面的请求到接收到该页面的所有数据的时间长度为服务响应时间,实际结果如图1所示)的结果进行预测模型建模。从测量结果数据中取1000个连续测量的结果作为样本。通过模型识别后采用 AR(3)预测模型进行预测,并与实际结果进行比较,结果如图2所示。预测的平均相对误差为10.7%。

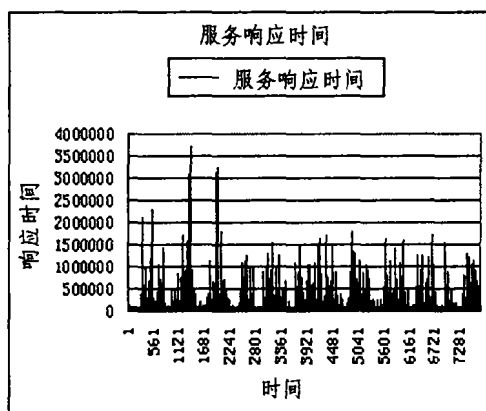


图1 S_{http} 的实际服务质量水平

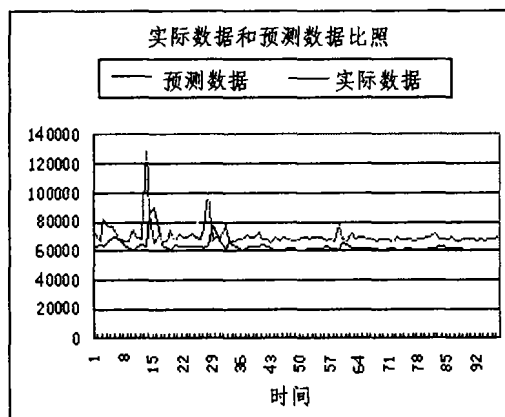


图2 连续采样数据应用预测模型结果

为了方便对照,其他采样方法是在连续采样获得的已有数据的基础上进行模拟采样实验的。选择服务的响应时间 SRT 作为自适应算法的参数,取质量条件 $\{SRT > 80000\text{ms}\}$ 作为违例条件,准确常数 $\alpha = 0.1$,采样周期为10分钟,初始采样期取100个周期进行准确函数的迭代计算。

表1给出了在相同采样周期的条件下使用系统采样方法、本文自适应采样方法和实际质量结果的对照情况。

表1 不同采样方法效果对照

采样方法	采样次数 ^{注1}	检测的违例次数	违例时长	时长偏差
服务连续采样	6500	3051	3051	—
系统采样	1300	607	3035 ^{注2}	0.52%
自适应采样	783	624	3040 ^{注3}	0.36%

注1:采样次数不包含初始采样期的迭代次数。注2:违例时长=检测的违例次数×采样区间长度。注3:违例时长=所有检测到违例的时间点到下次发生采样的时间点的距离之和。

可以看到,自适应方法不仅大大降低了采样需要的次数,而且甚至超过了其他传统方法(系统采样等)达到的精度。

表2 自适应算法的预测情况统计

采样方法	实际采样次数	由于预测而不采样的次数	比实际需要采样减少的比例	预测数据的平均误差
自适应采样	783	517	39.8%	13%

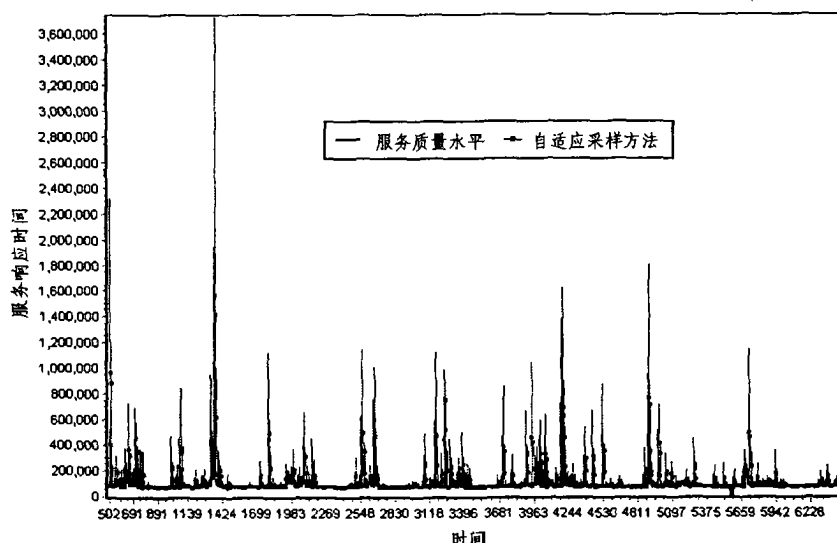


图3 自适应采样和实际服务质量的拟合情况

致谢 本文研究过程中,和软件开发环境国家重点实验室的郑庆国博士、徐莹博士进行了讨论,康建初教授和董卫宇博士也提出了一些很好的建议。

参考文献

- 1 梁静,石冰心. 服务水平协议的现状及未来. 电信快报, 2001-9
- 2 Matthews W, Cottrell L, Salomoni D. Passive and Active Monitoring on a High Performance Research Network. Passive and Active Monitoring (PAM) 2001
- 3 Zseby T, Molina M, Raspall F. Sampling and Filtering Techniques for IP Packet Selection. <http://www.ietf.org/drafts/draft-ietf-psamp-sample-tech-00.txt>, 2002
- 4 Amer P D, Cassel L N. Management of Sampled Real-Time Network Measurements. In: 14th Conf. on Local Computer Networks,

表2中可以看到,自适应算法减少了40%的采样次数,而且未采样的预测数据平均相对误差为13%,大部分数据基本和实际数据相吻合,是一个较好的结果。

图3显示了自适应采样结果和实际服务质量的拟合情况。

结论 本文针对服务质量水平的测量方法进行了分析,给出在给定精度要求和可信概率水平下采样次数的下界。

本文提出的基于自回归预测的自适应采样算法,将采样区间均分成小的时间区间,在每个区间进行随机采样。在进行采样时,使用自回归方法进行预测,并以一定的概率接受这个预测,从而减少了采样的次数。实验结果表明:该算法不但大大减少了采样的次数,而且预测精度在10%左右,在质量违例检测方面也比系统采样方法的效率高一些。

在本文算法中,特别重视了质量违例的检测,在预测结果满足违例条件时,以1的概率进行采样;如果不对质量违例情况加以区分的话,该采样方法需要采样的次数可以减少到50%以下。

本文提出的方法的关键是在服务采样中引入“预测”机制的思想,动态地构建预测模型和参数,提高模型的适应能力,进一步改善采样的效率和精度。必须注意的是,AR模型显然不会在所有的服务类型和参数中都如此有效,但是随着对各种服务的预测方法的研究,精度越高和效率越高的预测方法会不断地提出,在应用的过程中就可以不断地引入。

Minneapolis, IEEE, 1989. 62~68

- 5 Claffy K C, Polyzos G C, Braun H-W. Application of Sampling Methodologies to Network Traffic Characterization. In: Proc. of ACM SIGCOMM'93, San Francisco, CA, USA, Sep. 1993
- 6 Padmanabhan V N, et al. Server-based Inference of Internet Performance, [Microsoft Research Technical Report MSR-TR-2002-39]. May 2002
- 7 Choi B Y, Park J, Zhang Z-L. Adaptive random sampling for load change detection, [Technical Report]. University of Minnesota, Nov. 2001
- 8 Zseby T. Deployment of Sampling Methods for SLA Validation with Non-Intrusive Measurements. In: Proc. of Passive and Active Measurement Workshop (PAM 2002), Fort Collins, CO, USA, March, 2002