



基于大数据分析的网络异常检测方法

曹 旭,曹瑞彤

(中国联合网络通信集团有限公司 北京 100033)

摘 要:随着当前网络结构的日益复杂,网络故障出现的概率逐渐增加,网络管理面临更大的挑战。网络管理的一个重要任务是通过网络设备运行状况的监测,及时发现网络异常,提前进行处置,以降低故障带来的损失。当前业界已经对故障诊断进行了大量研究,但仍然存在告警丢失、虚警严重等问题。为此,提出了一种基于大数据分析的网络性能预测方法,通过改变网络管理中“先报警,再处理”的故障处理流程,实现网络故障的预警功能。本文的研究内容对于提升网络的可用性和可靠性具有一定的意义。

关键词:网络管理;故障预测;大数据;时间序列分析

doi: 10.3969/j.issn.1000-0801.2014.06.025

Network Anomaly Prediction Method Based on Big Data

Cao Xu, Cao Ruitong

(China United Network Communications Co., Ltd., Beijing 100033, China)

Abstract: With the development of the network technology and expansion of the network size, the probability of network anomaly is greatly increased, which makes network management face greater challenges. An important task of network management is network anomaly detection in time based on the data got by network management system (NMS); the performance of network management will be great improved if the network anomaly could be detected before it really happen. A method of network anomaly prediction based on big data was presented, the traditional process act as “first alarm, then solve” was replaced by anomaly prediction process. The result is significant for improving network availability and reliability.

Key words: network management, anomaly prediction, big data, time series analysis

1 引言

随着通信技术的飞速发展,电信网络在人们工作和生活中的作用越来越大,并从根本上改变了人们工作和生活的方式;同时,随着电信业务的不断深入发展,人们对网络的依赖性也不断增大。然而,日益增长的网络规模、不断更新的网络设备以及不断出现的新业务,使得网络出现故障的可能性大大增加,这给网络管理工作带来了全新的挑战。

网络管理的目的是通过对网络设备运行状况的连续监测,及时发现网络中的异常情况,并对异常情况及时报警,提醒网管人员采取措施保证网络正常运行。然而,在实际网络中,凭借网络管理人员的经验来分析网络运行状况的方式已经越来越难以满足运营商的需求。随着网络规模的扩大,网管人员常常需要面对大量告警而无法确定网络的真实故障。因此,迫切需要一种能够对网络数据进行智能处理的技术,即网络数据刚出现异常时,便能够自动检测出来,使故障在真正发生之前被发

现,从而可以大大提升网络管理人员的效率和网络的总体性能。

在传统的网络管理过程中,只有当网络设备出现异常并上报告警后,才会由网络管理人员解决问题,这显然是一种“响应式”的处理过程,即“先出现问题,再进行处理”的网络管理方式。然而,由于网络管理人员看到告警后,需要一段时间来分析确定问题才能采取措施,这种长时间的故障处理过程可能给用户和企业造成巨大的损失。

在网络运营的实践中,一些网络设备的故障和网络性能问题在发生之前就会在网络流量或性能参数中有所表现,但由于这些信息淹没在大量的数据中而没有引起网络管理者的重视。因此,通过大数据的方式挖掘和分析网络故障先兆,将使得网管人员有更多的时间分析问题、采取措施,避免出现更严重的故障,降低运营商和用户的损失。以此为出发点,提出了一种基于大数据思想发掘网络数据内部细节信息的方案,可实现对网络异常的及时发现,从而更大程度地提升网络可靠性。

2 网络故障预测的背景和模型

网络故障监测是网络管理的一项基本任务,也是网络管理人员的重要职责。网络故障监测的目的是提高网络的可靠性,在用户报告问题之前诊断或解决网络中的问题并提供网络规划和网络改造建议等。网络故障监测一般通过对网络流量、设备性能等网络参数的监测来判断网络状态和网络的运行情况。由于网络故障或网络性能问题通常会造成巨大的损失,基于网络参数实现准确、可靠的网络故障预测已经成为运营商迫切希望解决的问题。

基于网络故障的因果关系进行故障发现是一种直观的分析方法,包括IBM公司在内的大量企业和研究机构都提出了基于这一思路的网络故障发现方案^[1-2],但这种方案需要对网络故障成因和传播模式进行深入了解并建立模型。对于当前日益复杂的网络,采用这种方案所需的资金和时间成本都很难让运营商接受。

近年来,随着大数据技术的发展,大量基于观测数据的故障预测和分析方法已经被广泛运用到产品检测、生产过程控制、质量控制等领域,这类分析方法无需事先对故障的产生和传播建立复杂的模型,而只要通过基于统计学的分析就可以比较准确地发现故障,本文将这些领域的研究

成果引入网络管理领域,以更好地提升网络故障监测的性能表现。

在基于观测数据的网络故障方案中,最简单和普遍使用的是基于阈值的监测方法,即针对某个网络参数预先设定一个阈值,故障监测的任务就是监测该参数是否超过了这一阈值。这种方法简单易行,是实际网络管理中应用最广泛的方法,然而这种方法也存在两个明显的问题:一方面,阈值的选取是凭借网络管理人员的经验确定的,其取值缺乏可靠的依据,存在着较大的不确定性;另一方面,由于缺乏对阈值以下数据的处理,忽略了大量和网络状态有关的细节信息,如图1所示。

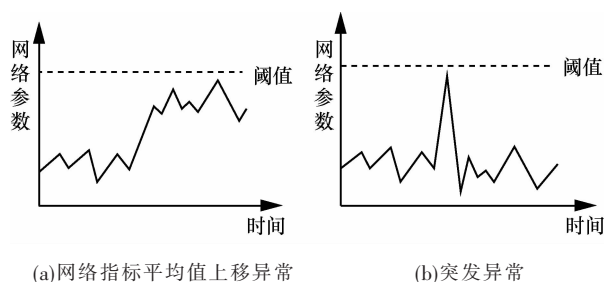


图1 阈值方法忽略网络状态细节信息

图1(a)显示了一种网络指标平均值上移的异常,这类情况的出现表明网络中出现了某类配置或业务异常,此时虽然网络指标没有超过阈值,但随着这一趋势的不断持续,有极大可能引发网络故障,因此应该提前进行处置。图1(b)则显示了一种突发异常,网络参数突然的大幅度变化预示了网络中的某种异常,如设备和网络过载、线路异常等情况,这种情况的发生也需要引起网络管理人员的注意。

为了弥补阈值方法的不足,尽可能地利用监测到的网络性能数据中的重要信息,需要加深对于网络“异常”的研究。在这方面,卡耐基梅隆大学的Maxion R A给出了网络状态“正常”和“异常”的定义。“正常”意味着符合常规或典型的模式,而“异常”则意味着违反了这种规律,与期望的情形有一定的偏差。

为了对网络状态进行定量分析,引入统计学理论,通过网络参数的变化对网络状态进行分析。与此同时,对网络参数的变化进行如下假设。

- 在一定的时间范围内,所监测到的网络参数是平稳的,即时刻 t 监测到的网络参数集 X_t 和时刻 $t+1$ 监测到的网络参数集 X_{t+1} 有相同的分布。



· 网络参数满足大数定律,即网络参数每天的监测

值收敛于期望值, $E(X) = \frac{1}{T} \sum_i X_i$ 。

基于上述假设,给出网络异常的统计学描述:网络参数的统计量出现了“明显”的偏差,也就是说,网络异常必然会引起网络参数在统计量上与正常情况的偏差。因此,发现异常的关键就是定量地对网络参数偏差的“显著性”进行分析。在第3节,本文将针对上述问题进行更加深入的讨论和分析。

3 基于网络行为模型的故障预测方法

如前文所述,当前的网络管理系统普遍可以收集大量的网络信息,而如何从大量的信息中自动获取维护网络正常运行的关键信息至关重要。网络参数异常意味着即将或已经发生某种网络或设备故障。在网络管理系统中,网络参数会根据时间的先后顺序进行存储,形成网络参数的时间序列,本文基于网络管理系统的这一特点,通过引入时间序列分析的方法展开网络故障预测的研究。

时间序列分析是一种基于动态数据发掘内在规律的统计分析方法,属于统计学科的一个分支。其思想是根据有限长度的历史记录建立精确反映系统内在运行规律的数学模型。在监测获得的网络参数的时间序列中,可以认为某一时刻 t 的网络参数集 X_t 由其之前若干时刻的数据以及随机干扰项所确定,即网络参数表现出一定的记忆特性。因此,可以建立 p 阶自回归 $AR(p)$ 模型:

$$X_t = \omega_1 X_{t-1} + \omega_2 X_{t-2} + \cdots + \omega_p X_{t-p} + e_t \quad (1)$$

其中, e_t 是白噪声, $\omega_1, \omega_2, \dots, \omega_p$ 表示此前各对应时刻对当前时刻的影响强度。式(1)表示的网络参数集带有 p 阶的记忆性,当网络状态发生变化或者产生“异常”时,上述关系会出现较大的偏差,具体体现为 e_t 的变化。因此,对于 e_t 的跟踪和分析是判断网络状态、预测网络异常的关键所在。

上述任务采用统计学方式描述:可以视为一个假设检验的问题,即设 $e_1, e_2, \dots, e_t$ 服从正态分布 H , 其均值和方差为 μ, σ^2 , 设 $E(e_{t+n}) = \mu + w$, 则网络异常需要判断 e_{t+n} 是否与分布 H 无差,即 w 是否为 0。 $e_1, e_2, \dots, e_t$ 的概率密度函数(PDF)可以写作:

$$f(\mu, \sigma^2, e_1, e_2, \dots, e_t) = \frac{1}{2(\pi\sigma^2)^{t/2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^t (e_i - \mu)^2 \right\} \quad (2)$$

因此, μ, σ^2 的最大似然估计是:

$$\bar{\mu} = \mu, \quad \bar{\sigma}^2 = \frac{1}{t+1} \sum_{i=1}^t (e_i - \mu)^2 \quad (3)$$

若网络可能存在异常,则对于 e_{t+n} , 其 μ, σ^2 应该写为:

$$\bar{\mu} = \frac{1}{n+1} (w + \sum_{i=1}^n e_i) = \frac{w}{n+1}, \quad \bar{\sigma}^2 = \frac{1}{n+1} \sum_{i=1}^n (e_i - \mu)^2 \quad (4)$$

于是,采用似然比检验的假设检验方式可以计算网络存在异常的似然比,似然比为:

$$\begin{aligned} \lambda = \frac{f_{H_1}}{f_{H_0}} &= \left[\frac{\sum_{i=1}^n [(e_i - \mu)^2 + (\mu - \bar{\mu})^2]}{\sum_{i=1}^t (e_i - \mu)^2} \right]^{n+1/2} \\ &= \left[1 + \left(\frac{\mu - \bar{\mu}}{\sigma} \right)^2 \right]^{n+1/2} \\ &= \left[1 + \left(\frac{1}{n+1} \right)^2 \left(\frac{w}{\sigma} \right)^2 \right]^{n+1/2} \end{aligned} \quad (5)$$

因此,可以通过不断对似然比 λ 进行监测获得对网络状态的判断。该方法的优点是所需数据量较少,计算复杂度较低,可以在保障一定可靠性的情况下快速对网络当前的状况进行分析判断,并对于网络可能出现的故障进行预先判断。基于该方法对网络管理系统进行升级,可以在充分利用运营商现有系统的前提下大大提升网络故障处理的效率。

4 方案实现与性能验证

如前所述,利用对网络参数进行时间序列分析可以更好地提升网络管理人员对于网络异常的发现效率。网络异常预警系统从对网络设备的数据采集开始,通过不断对采集到的数据进行时间序列分析,就可能在故障发生前实现对故障的提前预测,网络异常预警系统的系统架构如图2所示。

在对网络进行异常预警时,一个关键的步骤是确定模型的参数。以某骨干网络节点的网络流量数据为例,说明系统的实现原理。网络正常情况下在某个网络节点监测到的网络流量数据变化情况如图3所示。网络异常预警系统的任务就是首先建立网络正常情况的模型,发现网络参数偏离这一模型后,则产生异常报警。

采用 AR 模型对网络数据流量进行建模,本例中拟采用 $AR(2)$ 模型对数据流量进行建模。根据统计学相关理论, $AR(2)$ 模型满足平稳性的要求是:

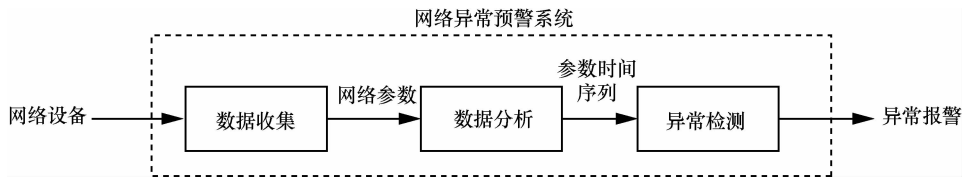


图 2 网络异常预警系统架构

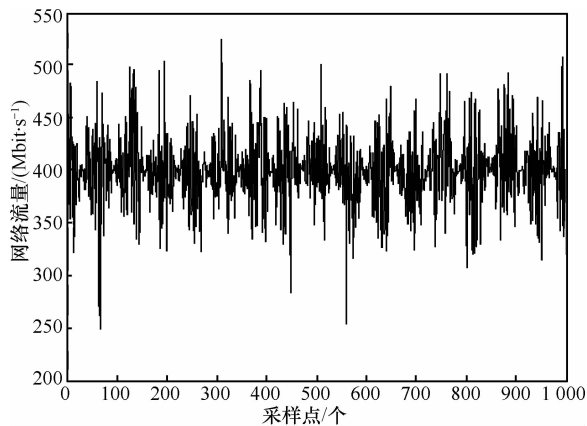


图 3 网络某节点的流量监测数据

$$\begin{cases} \omega_1 + \omega_2 < 1 \\ \omega_1 - \omega_2 < 1 \\ |\omega_2| < 1 \end{cases} \quad (6)$$

经过 $AR(2)$ 模型拟合,对于上述观测数据的 $AR(2)$ 模拟结果可以满足平稳性要求。如上文所述,对观测数据采用 $AR(2)$ 模型拟合后的残差 e_t 是网络故障分析的关键,观测数据经过处理后的残差结果如图 4 所示。

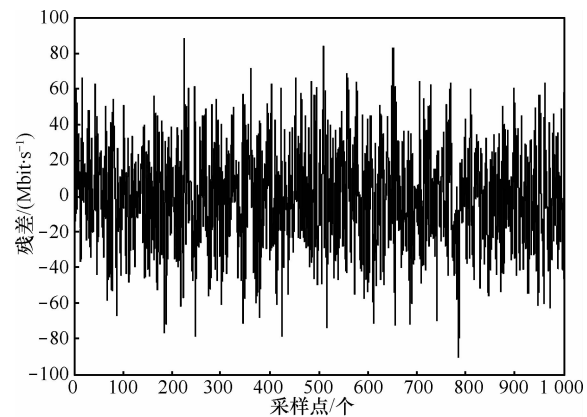


图 4 对样本采用 $AR(2)$ 模型拟合后的残差

为了更进一步验证模型的准确性,采用统计学中的 Q-Q 图 (quintile-quintile plot) 对残差进行分析。若网络不存在异常,则采集到的数据与模型相符合,计算出来的残差符合正态分布,此时残差结果在 Q-Q 图上呈线性;反之,则不成立。将图 4 的结果用 Q-Q 图进行呈现,如图 5 所示。

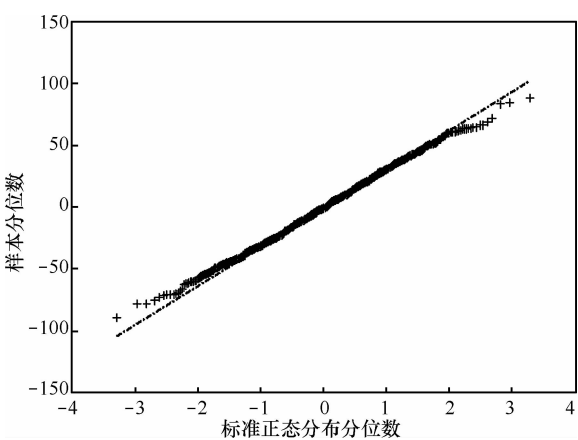


图 5 样本残差的 Q-Q 图

从图 5 中可以看到,样本经过处理后的残差呈现出较强的线性特点,基本服从正态分布,因此采样数据不含网络异常数据。这一结果也从一个方面证明了本文所提方法的有效性。与此类似,网络的异常检测则是基于对采样数据残差偏离正态分布的情况实现的,由于其判别过程与上述流程类似,且异常产生的表现多种多样,本文不再累述。

实际上,运营商对网络异常预警的需求不仅是尽可能准确地检查出网络中可能的故障,同时还要求对于非故障错误判断的虚警率尽可能小。对于网络故障预警系统的性能判断也应该全面考虑上述两点需求。本文以某骨干网一周的监测和报警的历史数据为基础,将历史数据输入本文所提出的系统来验证该方案的可行性,其结果判断通过对比告警处理的工单来确定。对结果进行统计分析,并通过 ROC 曲线 (receiver operating characteristic curve) 进行呈现,结果如图 6 所示。

从图 6 可以看到,本文所提出的故障预警方案不仅故障判断总体正确率明显高于传统的阈值方案,同时虚警率的表现也好于传统的阈值方案。虽然,采用本文所提算法在一定程度上提升了系统的计算复杂度,然而试验表明,采用本文所提方案可以将故障检测准确率提升到 90% 以上。这种性能提升显示了本文所提方案具有一定的使用价值,对该算法进行进一步改进和验证将是笔者未来工作的重要方向。

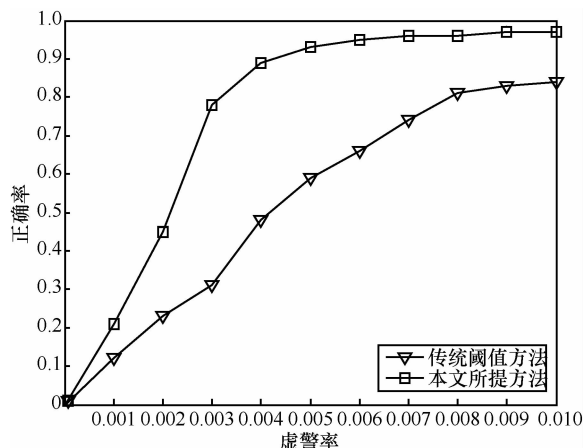


图6 两种方法的 ROC 曲线对比

5 结束语

随着网络规模的扩大和通信技术的发展,网络管理工作也面临着越来越多的挑战。本文就网络故障的预测和监测方法进行了研究,希望基于大数据的思想,充分挖掘蕴含在大量网络数据中的有用信息,并将之用于网络故障发现和预测。从本文的结论可以看到,大数据和网络管理方法的结合已经显现出越来越大的潜力,随着两者结合的不断深入,必将为未来网络管理的发展带来更多的突破。

参考文献

- 1 Hanemann A, Sailer M, Schmitz D. Towards a framework for it service fault management. Proceedings of the European

University Information Systems Conference (EUNIS2005), Manchester, England, 2010

- 2 Steindler M, Sethi A S. Probabilities fault diagnosis in communication systems through incremental hypothesis updating. Computer Networks, 2011, 45(4):537~562
- 3 Box E P, Jenkins G M, Reinsel G C. 时间序列分析—预测与控制. 顾岚, 范金减译. 北京: 中国统计出版社, 2011
- 4 Basu S, Mukherjee A, Klivansky S. Time series models for internet traffic. <http://hdl.handle.net/1853/6696>, 1996
- 5 Frost V, Melamed B. Traffic modeling for telecommunications networks. IEEE Communication Magazine, 2004, 32(3):70~81

[作者简介]



曹旭,男,中国联合网络通信集团有限公司中级工程师,主要研究方向为计算机网络技术、IT 信息化运维管理。



曹瑞彤,女,中国联合网络通信集团有限公司中级经济师,主要研究方向为通信技术及通信经济发展。

(收稿日期:2014-06-08)

(上接第 151 页)

局视图的 PCE 或 SDN 集中控制器,可以方便地实现光层的重路由。

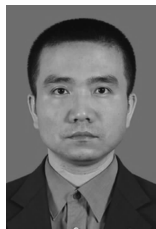
6 结束语

信号处理的最低层面是光层(L0 层),光层上面是电层(L1 层),一般来说光层比电层的处理成本、功耗都要低,而效率却更高;由于直接的光层处理对波长信号的速率是透明的,非常适合应用在未来单波 100 Gbit/s 以上的超大容量波分传输系统中,随着互联网和云计算技术的发展,云数据中心之间直连的需求会越来越多,带宽也将越来越大,直接在光层传送超大容量的信号将是一种理想的选择,未来具备 CDC 功能的 ROADM 将得到广泛的应用。

参考文献

- 1 周斌,张阳安,黄永清等. 基于波长选择开关的 ROADM 实现研究. 光通信研究, 2008, 145(1): 11~13
- 2 李俊杰. ROADM 技术的应用. 中兴通讯技术, 2013, 19(3): 26~30

[作者简介]



赵春华,男,博士,广东省电信规划设计院有限公司高级工程师、副总工程师、注册咨询工程师,长期从事传送网、IP 网的咨询规划设计和技术管理工作。

(收稿日期:2014-04-16)