

# 基于机器学习的智能运维

裴丹<sup>1</sup> 张圣林<sup>2</sup> 裴昶华<sup>3</sup>

<sup>1</sup> 清华大学

<sup>2</sup> 南开大学

<sup>3</sup> 阿里巴巴公司

关键词：机器学习 智能运维

当代社会生产生活的许多方面都依赖于大型复杂的软硬件系统，包括互联网、高性能计算、电信、金融、电力网络、物联网、医疗网络和设备、航空航天、军用设备及网络等。这些系统的用户都期待有好的体验。因而，这些复杂系统的部署、运行和维护都需要专业的运维人员，以应对各种突发事件，确保系统安全、可靠地运行。由于各类突发事件会产生海量数据，因此，智能运维从本质上可以认为是一个大数据分析的具体场景。

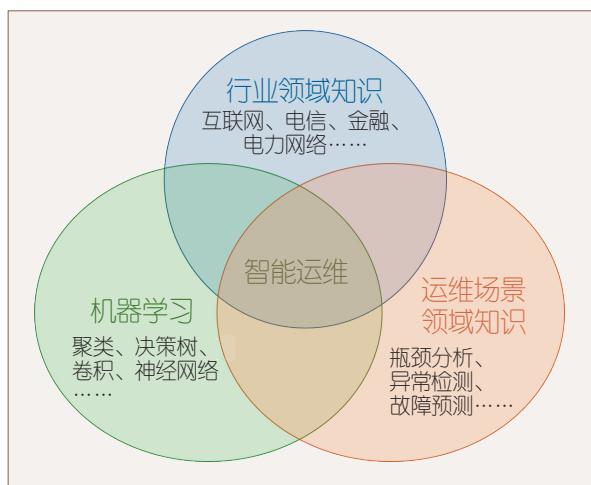


图1 智能运维涉及的范围

图1展示了智能运维涉及的范围。它是人工智能、行业领域知识、运维场景领域知识三者相结合的交叉领域，离不开三者的紧密合作。

## 智能运维的历史

**手工运维：**早期的运维工作大部分是由运维人员手工完成的，那时，运维人员又被称为系统管理员或网管。他们负责的工作包括监控产品运行状态和性能指标、产品上线、变更服务等。因此，单个运维人员的工作量和运维人员的数量都是随着产品的个数或者产品服务的用户规模呈线性增长的。此时的运维工作消耗大量的人力资源，但大部分运维工作都是低效的重复。这种手工运维的方式必然无法满足互联网产品日新月异的需求和突飞猛进的规模。

**自动化运维：**运维人员逐渐发现，一些常见的重复性的运维工作可以通过自动化的脚本来实现：一部分自动化脚本用以监控分布式系统，产生大量的日志；另外一部分被用于在人工的监督下进行自动化处理。这些脚本能够被重复调用和自动触发，并在一定程度上防止人工的误操作，从而极大地减少人力成本，提高运维的效率。这就诞生了自动化运维。自动化运维可以认为是一种基于行业领域知识和运维场景领域知识的专家系统。

**运维开发一体化：**传统的运维体系将运维人员从产品开发人员中抽离出来，成立单独的运维部门。这种模式使得不同公司能够分享自动化运维的工具和想法，互相借鉴，从而极大地推动了运维的发展。然而，这种人为分割的最大问题是产生了两个对立的团队——产品开发人员和运维人员。他们

的使命从一开始就截然不同：产品开发人员的目标是尽快地实现系统的新功能并进行部署，从而让用户尽快地使用到新版本和新功能。运维人员则希望尽可能少地产生异常和故障。但是经过统计发现，大部分的异常或故障都是由于配置变更或软件升级导致的。因此，运维人员本能地排斥产品开发人员部署配置变更或软件升级。他们之间的目标冲突降低了系统整体的效率。此外，由于运维人员不了解产品的实现细节，因此他们在发现问题后不能很好地定位故障的根本原因。为了解决这一矛盾，DevOps<sup>1</sup>应运而生。DevOps最核心的概念是开发运维一体化，即不再硬性地区分开发人员和运维人员。开发人员自己在代码中设置监控点，产生监控数据。系统部署和运行过程中发生的异常由开发人员进行定位和分析。这种组织方式的优势非常明显：能够产生更加有效的监控数据，方便后期运维；同时，运维人员也是开发人员，出现问题之后能够快速地找出根因。谷歌的站点可靠性工程（Site Reliability Engineering, SRE）就是DevOps的一种特例。

**智能运维（Artificial Intelligence for IT Operations, AIOps）：**自动化运维在手动运维基础上大大提高了运维的效率，DevOps有效地提升了研发和运维的配合效率。但是，随着整个互联网系统数据规模的急剧膨胀，以及服务类型的复杂多样，“基于人为指定规则”的专家系统逐渐变得力不从心。因为自动化运维的瓶颈在于人脑：必须由一个长期在一个行业从事运维的专家手动地将重复出现的、有迹可循的现象总结出来，形成规则，才能完成自动化运维。然而，越来越多的场景表明，简单的、基于人为制定规则的方法并不能够解决大规模运维的问题。

与自动化运维依赖人工生成规则不同，智能运维强调由机器学习算法自动地从海量运维数据（包括事件本身以及运维人员的人工处理日志）中不断地学习，不断地提炼并总结规则。换句话说，智能运维在自动化运维的基础上增加了一个基于机器学

习的大脑，指挥着监测系统采集大脑决策所需的数据，做出分析、决策并指挥自动化脚本去执行大脑的决策，从而达到运维系统的整体目标（见图2）。Gartner Report 预测 AIOps 的全球部署率将从 2017 年的 10% 增加到 2020 年的 50%。

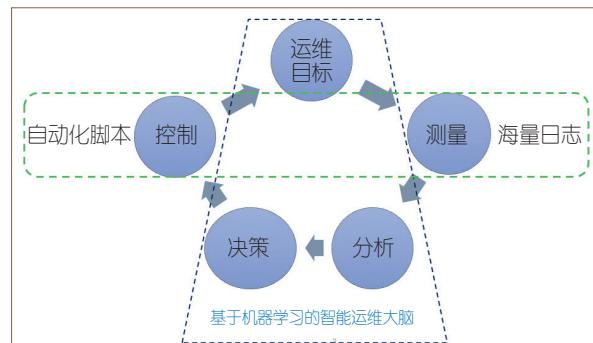


图2 智能运维与自动化运维的最大区别是有一个基于机器学习的大脑

## 智能运维现状

### 关键场景与技术

图3显示了智能运维包含的关键场景和技术，涉及大型分布式系统监控、分析、决策等。

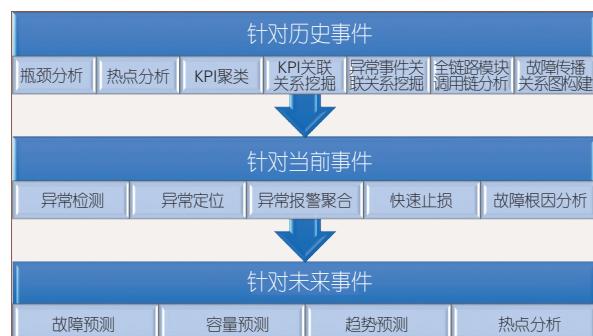


图3 智能运维的关键场景和技术

在针对历史事件的智能运维技术中，**瓶颈分析**是指发现制约互联网服务性能的硬件或软件瓶颈。**热点分析**指的是找到对于某项指标（如处理服务请

<sup>1</sup> DevOps，英文 Development 和 Operations 的组合。是一组过程、方法与系统的统称，用于促进开发（应用程序 / 软件工程）、技术运营和质量保障 (QA) 部门之间的沟通、协调与整合。

求规模、出错日志)显著大于处于类似属性空间内其他设施的集群、网络设备、服务器等设施。**KPI<sup>2</sup>曲线聚类**是指对形状类似的曲线进行聚类。**KPI曲线关联挖掘**针对两条曲线的变化趋势进行关联关系挖掘。**KPI曲线与报警之间的关联关系挖掘**是针对一条KPI曲线的变化趋势与某种异常之间的关联关系进行挖掘。**异常事件关联挖掘**是指对异常事件之间进行关联关系挖掘。**全链路模块调用链分析**能够分析出软件模块之间的调用关系。**故障传播关系图构建**融合了上述后四种技术,推断出异常事件之间的故障传播关系,并作为故障根因分析的基础,解决微服务时代KPI异常之间的故障传播关系不断变化而无法通过先验知识静态设定的问题。通过以上技术,智能运维系统能够准确地复现并诊断历史事件。

针对当前事件,**KPI异常检测**是指通过分析KPI曲线,发现互联网服务的软硬件中的异常行为,如访问延迟增大、网络设备故障、访问用户急剧减少等。**异常定位**在KPI被检测出异常之后被触发,在多维属性空间中快速定位导致异常的属性组合。**快速止损**是指对以往常见故障引发的异常报警建立“指纹”系统,用于快速比对新发生故障时的指纹,从而判断故障类型以便快速止损。**异常报警聚合**指的是根据异常报警的空间和时间特征,对它们进行聚类,并把聚类结果发送给运维人员,从而减少运维人员处理异常报警的工作负担。**故障根因分析**是指根据故障传播图快速找到当前应用服务KPI异常的根本触发原因。故障根因分析系统找出异常事件可能的根因以及故障传播链后,运维专家可以对根因分析的结果进行确定和标记,从而帮助机器学习方法更好地学习领域知识。这一系统最终达到的效果是当故障发生时,系统自动准确地推荐出故障根因,指导运维人员去修复或者系统自动采取修复措施。

## 关键技术示例

### KPI瓶颈分析

为了保证向千万级甚至上亿级用户提供可靠、

高效的服务,互联网服务的运维人员通常会使用一些关键性能指标来监测这些应用的服务性能。比如,一个应用服务在单位时间内被访问的次数(Page Views, PV),单位时间交易量,应用性能和可靠性等。KPI瓶颈分析的目标是在KPI不理想时分析系统的瓶颈。一般监控数据中的关键指标有很多属性,这些属性可能影响到关键指标,如图4所示。



图4 KPI及影响因素

当数据规模较小时,运维人员通过手动过滤和选择,便能够发现影响关键性能指标的属性组合。但是,当某个关键指标有十几个属性,每个属性有几百亿条数据时,如何确定它们的属性是怎样影响关键性能指标的,是一个非常有挑战性的问题。显然,采用人工的方式去总结其中的规律是不可行的。因此,需要借助于机器学习算法来自动地挖掘数据背后的现象,定位系统的瓶颈。

针对这一问题,学术界已经提出了层次聚类、决策树、聚类树(CLTree)等方法。FOCUS<sup>[1]</sup>通过对数据预处理,把KPI分为“达标”和“不达标”两类,从而把KPI瓶颈分析问题转化为在多维属性空间中的有监督二分类问题。由于瓶颈分析问题要求结果具备可解释性,因此FOCUS采用了结果解释性较好的决策树算法。该算法较为通用,可以针对符合图4所示的各类数据进行瓶颈分析。

### KPI异常检测

**KPI异常检测**是互联网服务智能运维的一个底层核心技术。上述大多数智能运维的关键技术都依赖于KPI异常检测的结果。

当KPI呈现出异常(如突增、突降、抖动)时,

<sup>2</sup> 关键性能指标, Key Performance Indicator。

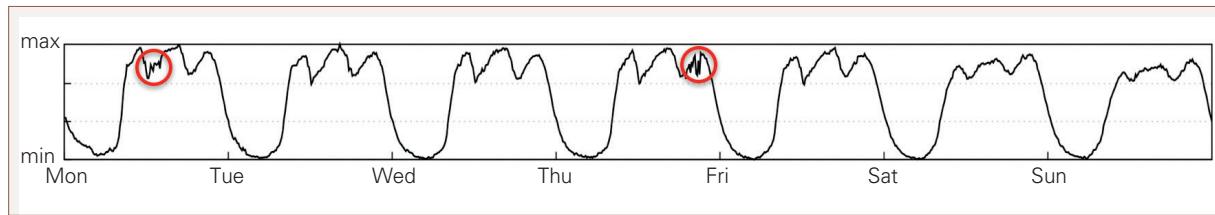


图5 KPI异常示例：某搜索引擎PV曲线的异常

往往意味着与其相关的应用发生了一些潜在的故障，比如网络故障、服务器故障、配置错误、缺陷版本上线、网络过载、服务器过载、外部攻击等。图5展示了某搜索引擎一周内的PV数据，其中红圈标注的为异常。

因此，为了提供高效、可靠的服务，必须实时监测KPI，以便及时发现异常。同时，那些持续时间相对较短的KPI抖动也必须被准确检测出来，以避免未来的经济损失。

目前，学术界和工业界已经提出了一系列KPI异常检测算法。这些算法可以概括地分成基于窗口的异常检测算法，例如奇异谱变换(singular spectrum transform)；基于近似性的异常检测算法；基于预测的异常检测算法，例如Holt-Winters方法、时序分解方法、线性回归方法、支持向量回归等；基于隐式马尔科夫模型的异常检测算法；基于分段的异常检测算法；基于机器学习(集成学习)的异常检测算法<sup>[2]</sup>等类别。

### 故障预测

现在，主动的异常管理已成为一种提高服务稳定性的有效方法。故障预测是主动异常管理的关键技术。故障预测是指在互联网服务运行时，使用多种模型或方法分析服务当前的状态，并基于历史经验判断近期是否会发生故障。

图6显示了故障预测的定义。在当前时刻，根据一段时间内的测量数据，预测未来某一时间区间是否会发生故障。之所以预测未来某一时间区间的故障，是因为运维人员需要一段时间来应对即将发生的故障，例如切换流量、替换设备等。

目前，学术界和工业界已经提出了大量的故障预测方法。大致可分为几个类别：

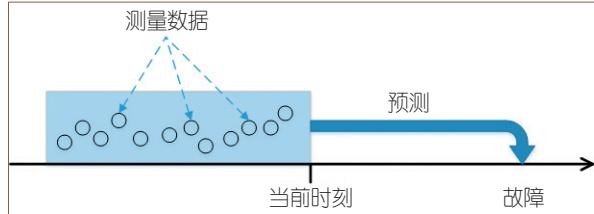


图6 故障预测定义

- 故障踪迹。其核心思想是从以往故障的发生特征上推断即将发生的故障。发生特征可以是故障的发生频率，也可以是故障的类型。
- 征兆监测。通过一些故障对系统的“副作用”来捕获它们，例如，异常的内存利用率、CPU使用率、磁盘I/O、系统中异常的功能调用等。
- 错误记录。错误事件日志往往是离散的分类数据，例如事件ID、组件ID、错误类型等。

### 智能运维所用到的机器学习算法

在智能运维文献中较为常见的算法包括逻辑回归、关联关系挖掘、聚类、决策树、随机森林、支持向量机、蒙特卡洛树搜索、隐式马尔科夫、多示例学习、迁移学习、卷积神经网络等。在处理运维工单和人机界面时，自然语言处理和对话机器人也被广泛应用。

智能运维系统在演进的过程中，不断采用越来越先进的机器学习算法。

基于互联网的视频流媒体（如QQ视频、优酷、爱奇艺、Netflix等）已经逐渐渗透到人们的日常生活中。在网络领域顶级会议中也涌现了很多学术界和工业界合作的智能运维案例，如卡内基梅隆大学的系列工作：SIGCOMM’11论文<sup>[3]</sup>利用不同数据分析及统计分析方法，灵活使用可视化(visualization)、

相关分析 (correlation)、信息熵增益 (information gain) 等工具，将杂乱无章的数据转化为直观清晰的信息，从而分析出海量数据背后的视频体验不佳的规律和瓶颈；SIGCOMM'12 论文<sup>[4]</sup> 为视频传输设计了一个“大脑”，根据视频客户和网络状况的全局信息，动态地优化视频传输；SIGCOMM'13 论文<sup>[5]</sup> 通过决策树模型建立视频流媒体用户参与度的预测模型，指导关键性能指标的优化策略，最终有效地改善了视频流媒体用户的体验质量；NSDI'17 论文<sup>[6]</sup> 将视频质量的实时优化问题转化为实时多臂老虎机问题（一种基础的强化学习方法），并使用上限置信区间算法有效解决了这一问题。这一系列论文，见证了智能运维的不断演进之路。

## 智能运维未来展望

多个行业领域都表现出对智能运维的强烈需求。但是，他们主要在各自行业内寻找解决方案。同时，受限于所处行业运维团队的开发能力，他们往往对所处行业内的运维团队提出相对较低的需求——这些需求一般停留在自动化运维的阶段。如果各行业领域能够在深入了解智能运维框架中关键技术的基础上，制定合适的智能运维目标，并投入适当的资源，一定能够有效地推动智能运维在各自行业的发展。

在基于机器学习的智能运维框架下，机器将成为运维人员的高效可靠的助手。运维工程师逐渐转型为大数据工程师，负责搭建大数据基础架构，开发和集成数据采集程序和自动化执行脚本，并实现高效的机器学习算法。同时，在面对所处行业的智能运维需求时，智能运维工程师可以在整个智能运维框架下跨行业地寻找关键技术，从而能够更好地满足本行业的智能运维需求，达到事半功倍的效果。

智能运维的基石是机器学习和人工智能。相比人工智能在其他领域的应用，智能运维几乎完美地拥有一个有前景的人工智能垂直应用领域必备的要素：实际应用场景、大量数据、大量标注。智能运维几乎所有的关键技术都离不开机器学习算法；工业界不断产生海量运维日志；由于运维人员自身就

是领域专家，其日常工作就会产生大量的标注数据。海量的数据和标注降低了研究机器学习算法的门槛，有益于算法研究快速取得进展。因此，智能运维可以说是机器学习领域一个尚未开采的“金矿”，非常值得机器学习领域科研人员的关注和投入。

作为人工智能的一个垂直方向，智能运维的理论也将取得长足的进步。除了互联网以外，智能运维在高性能计算、电信、金融、电力网络、物联网、医疗网络和设备、航空航天、军用设备及网络都将有很好的应用。



裴丹

CCF 专业会员。清华大学计算机系长聘副教授，特别研究员，青年千人，ACM/IEEE 高级会员。主要研究方向为基于机器学习的智能运维。在智能运维领域发表了 90 余篇学术论文，20 多项美国专利授权。peidan@tsinghua.edu.cn



张圣林

CCF 专业会员，CCF 互联网专业委员会委员。南开大学软件学院讲师。主要研究方向为网络管理、数据挖掘、数据中心。zhangsl@nankai.edu.cn



裴旭华

CCF 专业会员。阿里巴巴公司猜你喜欢部门高级算法工程师。主要从事手机淘宝商品推荐相关的研究工作。changhua.pch@alibaba-inc.com

## 参考文献

- [1] Liu D, Zhao Y, Sui K, et al. FOCUS: Shedding Light on the High Search Response Time in the Wild[C]// Proceedings of INFOCOM 2016, 2016:1-9.
- [2] Liu D, Zhao Y, Xu H, et al. Oppentice: Towards Practical and Automatic Anomaly Detection Through Machine Learning[C]// Proceedings of the 2015 Internet Measurement Conference. New York: ACM Press, 2015:211-224.
- [3] Dobrian F, Sekar V, Awan A, et al. Understanding the impact of video quality on user engagement[C]// ACM SIGCOMM Computer Communication Review. ACM, 2011, 41(4): 362-373.
- [4] Liu X, Dobrian F, Milner H, et al. A case for a

coordinated internet video control plane[C]//*Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication*. ACM, 2012: 359-370.

[5] Balachandran A, Sekar V, Akella A, et al. Developing a predictive model of quality of experience for internet

video[C]//*ACM SIGCOMM Computer Communication Review*. ACM, 2013, 43(4): 339-350.

[6] Jiang J, Sun S, Sekar V, et al. Pytheas: Enabling Data-Driven Quality of Experience Optimization Using Group-Based Exploration-Exploitation[C]//*NSDI*. 2017: 393-406.