# 智能运维在中国移动 IT 云中的应用与实践

刘　虹，滕　滨，张　琳，郭志斌
（中国移动通信集团有限公司 信息技术中心，北京 100032）

**摘　要：** 介绍了中国移动 IT 云针对 IaaS 层的智能运维场景体系规划，选择了数据基础较好的两个典型应用场景“智能化指标异常检测”和“智能化告警关联与溯源”进行了研究和论证，分别分析了两个场景适用的算法和实现过程，论述了两个场景实施后的效果评估方法，并经实际生产验证了场景实施的效果。

**关键词：** 智能运维；指标异常检测；告警溯源；机器学习
**中图分类号：** TN929.5；TP399　　**文献标识码：** A　　**DOI：**10.16157/j.issn.0258-7998.211543

**中文引用格式：** 刘虹，滕滨，张琳，等. 智能运维在中国移动 IT 云中的应用与实践[J]. 电子技术应用，2021，47(11)：20-24.
**英文引用格式：** Liu Hong，Teng Bin，Zhang Lin，et al. Best practice of AIops in China Mobile private cloud[J]. Application of Electronic Technique，2021，47(11)：20-24.

## Best practice of AIops in China Mobile private cloud

Liu Hong，Teng Bin，Zhang Lin，Guo Zhibin
（Information Technology Center，China Mobile Communications Group Co.，Ltd.，Beijing 100032，China）

**Abstract：** The planning of infrastructure AIops scenario for China Mobile private cloud is described，and the two typical scenarios named "Intelligent Index Anomaly Detection" and "Intelligent Alarm Traceability" are researched. The algorithm and business processes of the two scenarios are introduced respectively.The effect evaluation method of the two scenarios is discussed，and the actual production verifies the implementation effect.

**Key words：** AIops；index anomaly detection；alarm traceability；machine learning

## 0 引言

随着国内企业数智化转型的深入推进，企业私有云的设备规模呈现持续增加的趋势，作为中国移动内部支撑系统的云化基础设施，一级云资源池的规模持续增加，运营和运维工作面临着越来越大的压力。从业界经验来看，运维人员数量无法随着设备数量线性增加，每万台服务器运维人员的数量持续下降，因此亟需引入智能化运维手段，解决人力不足的矛盾。同时，也需要借助智能化工具提高资源的可用性，提升租户的使用体验。为此，中国移动结合 IT 云自身特点，梳理了一级 IT 云的智能运维场景体系，并选取典型场景进行了应用与实践。本文基于中国移动一级 IT 云运维团队的切实需求，综合评估业界关键技术成熟度和一级 IT 云的基础运维数据质量，选择以下两个场景进行分析和研究：

(1)智能化的指标异常检测

通过机器学习算法，从监控指标的历史数据中识别指标的特征，并基于指标特征生成指标的个性化异常检测模型[1]。本场景希望解决传统固定阈值的检测精度不足的问题，并缓解人工设置阈值的决策困境，降低工作量。

(2)智能化的告警关联与溯源

通过机器学习算法，从海量历史告警中学习告警之间的相关性，结合网络拓扑结构及专家标注，实现告警的智能关联压制(聚合)，并推断告警根源。本场景希望提升运维人员在面对故障引起的大量告警时，能够快速定位问题，提高故障恢复的速度。

## 1 智能化指标异常检测

面对一级 IT 云全网资源池设备 10 万+、指标数量千万级的实际情况(典型的服务器设备有近百个监控的指标，典型的网络设备则有数百个监控的指标)，事实上，已经无法依靠人工为每个设备的指标都配置合适的阈值。况且，设备上的监控指标，其波形因其承载的业务不同而千变万化，人工设置的固定阈值无法适应指标的动态性，容易产生误报或者漏报。

因此引入人工智能算法，实现智能化的指标异常检测的需求应运而生，该场景的总体工作流程如图 1 所示。本文着重介绍周期性指标的分析算法。

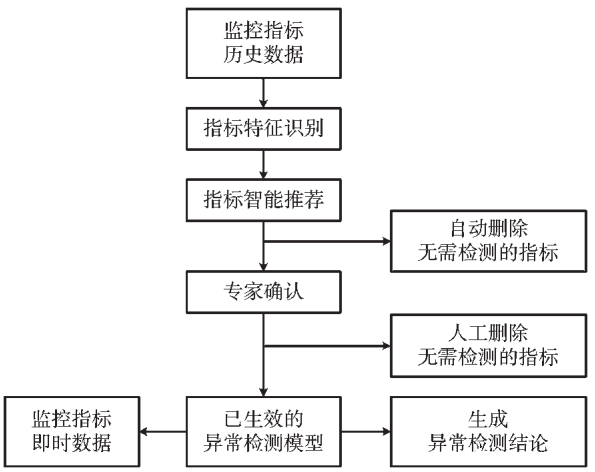



图 1　指标异常检测流程

### 1.1 周期型指标的特征识别

#### 1.1.1 数据预处理

原始的指标数据往往包含噪声，常见的有周期错位、数据缺失/极值等，会影响模型的训练。预处理负责对各种异常情况进行对应的处理，得到标准的、干净的、连续的数据，供给特征分析使用。

#### 1.1.2 离群点处理

离群点是明显背离指标分布的点，离群点检测主要采用 LOF 局部异常因子算法[2-3]，识别出离群点后，可以采用平均数填充、中数填充、重复值最多的数填充、丢弃等策略。

#### 1.1.3 数据转换

数据转换是根据指标特征将数据进行转换或归并，从而构成一个适合特定特征分析的描述形式。

数据转换主要包括 3 种策略：

(1)标准化，标准化给定数据集中所有数值属性的值到一个 0 均值和单位方差的正态分布；

(2)归一化，规范化给定数据集中的所有数值属性值，类属性除外，结果值默认在区间[0，1]；

(3)离散化，分别进行监督和无监督的数值属性的离散化，用来离散数据集中的一些数值属性到分类属性。

#### 1.1.4 识别周期性

首先对原始数据 $K=\{v_1,v_2,\cdots,v_n\}$ 做傅里叶变换，获取其振幅最大的分量作为备选周期 $T$。将数据按照备选周期 $T$ 进行切分，形成 $N$ 个子序列：$K_1$~$K_n$。

任选其中的一个子序列(如 $K_1$)作为基准，计算其与其他 $N-1$ 个子序列的皮尔逊系数，并求均值。皮尔逊系数用于度量两个变量 $X$ 和 $Y$ 之间的相关性，是非常成熟的算法，不再赘述[4-5]。

若 $K_n$ 的皮尔逊系数均值达不到参数配置的阈值要求，则轮换 $K_{n+1}$ 为基准，直到找出满足阀值要求的子序列。如果能够找到子序列 $K_m$ 满足要求，则判定该指标具有周期性特征，否则判定该指标不具备周期性特征。

#### 1.1.5 识别节假日效应

如文献[6]所述，时间序列数据应充分考虑假日效应，本文首先将上一步生成的子序列按照其所在日期是否为假日分为两组，记为 $C_w$ 和 $C_h$。然后，在 $C_w$ 和 $C_h$ 上分别运行主成分分析(PCA)[7-8]，各选定一个最具代表性的子序列，记为 $K_w$ 和 $K_h$，通过计算它们之间的皮尔逊系数来判定曲线的波形是否相似。如果两者波形非常相似，那么说明该指标不具备节假日效应，反之则具有节假日效应。

### 1.2 周期型指标的基线计算

经过数据预处理，可以得到一个(无假日效应)时或者两个(有假日效应时)基于历史数据的叠加图，将一天的数据按照采集周期分成 $N$ 个时间点，每个时间点相当于一个桶，将历史数据分别放到每个桶里，然后计算出每个桶的均值、最大值和最小值，即为该时刻的基本基线值。将每个时刻的最大值、最小值分别连接起来，就得到该模型周期的基本基线。直接计算得出的基线非常生硬且非常敏感，因此系统提供一些参数来降低基线的敏感度，防止造成告警误报。

### 1.3 周期型指标的异常检测

通过分析得到周期性指标的动态基线模型后，还需要对模型进行测试以验证本文的模型是否准确，如果不准确可以随时调整参数对模型进行微调。对于短暂的基线偏离，如果认为这个点的短暂偏离是正常现象，可以调整参数，以修改基线的敏感度。

### 1.4 应用效果

选择数据中心作为试点，进行“智能化的指标异常检测”场景相关工具的落地验证。

#### 1.4.1 特征识别可靠性评估

验证环节邀请了三位一线运维专家对指标特征进行评估，并与算法的智能识别结果进行对比，部分结果如表 1 所示。

表 1　指标特征识别与专家评估意见

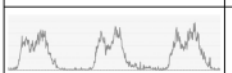
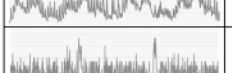
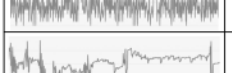
| 指标波形 | 专家意见 | | | 特征置信度 | | | 结论 |
|---|---|---|---|---|---|---|---|
| | 专家 1 | 专家 2 | 专家 3 | 周期 | 趋势 | 平稳 | |
|  | 周期 | 周期 | 周期 | 96 | 2 | 38 | 吻合 |
|  | 趋势 | 趋势 | 趋势 | 23 | 89 | 42 | 吻合 |
|  | 周期 | 周期 | 周期 | 98 | 1 | 66 | 吻合 |
|  | 无特征 | 周期 | 平稳 | 54 | 4 | 56 | 难以界定 |
|  | 平稳 | 平稳 | 平稳 | 24 | 40 | 85 | 吻合 |
|  | 无特征 | 无特征 | 平稳 | 15 | 11 | 49 | 难以界定 |

表 1 中特征置信度最高的值字体加下划线并加粗，代表系统自动选定的特征。其中周期特征指曲线以相对固定的周期重复类似的形态，趋势特征指曲线呈现递增或者递减的形态，平稳特征是指指标在一个箱体内随机振荡。

从表 1 中可以看到，当人工智能识别出的指标特征置信度大于 80%时，与专家的评估意见高度吻合(如第 1，2，3，5 行)；而当指标特征置信度小于 60%时(如第 4，6 行)，可以认为指标不具备明显特征。

1.4.2 效率提升情况

本次试验接入的 1 081 台设备，按照其中有 5%的网络设备来算，可得这些设备的指标总数有：(1 081×0.05)×500+(1 081×0.95)×80=27 025+82 156≈11 万(个)。

如果通过传统手工方式设置监控策略，即便每个指标耗费运维人员 1 min 的时间(该估计已经非常乐观)，则共需：110 000/60/8/22≈10.4(人月)。

作为对比，本文在 20 个设备上做了一次特征识别，选择了 2019 年 3 月份的数据作为训练集，并选择了 4 月第一周的数据作为测试集。这 20 个设备共有 3 169 个有效指标，特征识别结果的分布如图 2 所示。

系统自动推荐了 17 个设备上的 356 个特征置信度比较高的指标供运维人员做确认，占总指标数的 11.23%，系统运行耗时间为 10 min，可以忽略。由图 2 可见，通过引入智能分析，将指标梳理这项机械性工作的工作量降低了 90%左右，使得一项看似不可能完成的工作变成了可能。

## 2 智能化告警关联和溯源

在云计算环境中，业务系统或设备间存在各种依赖关系，因此在系统内或者系统间就会存在故障关联，也就是当系统中一个模块或者设备发生告警时，与之关联的模块或设备也往往发生告警[9-10]。

如文献[11]所述的案例法，在中国移动一级 IT 云资源池在 10 万+的设备级别上极难开展，需要基于历史告警数据学习告警之间的相关性，实现告警的智能压制、推断告警根源，有效提高告警有效性。

### 2.1 告警溯源分析流程

告警溯源分析工作流程的要点在于：

(1)在告警数据集上，基于告警数据的特点，运行多次的全量基础扫描，并基于基础扫描的结果做定点的深度扫描，从而发现不同告警的相关性；

(2)引入性能监控指标数据集，在性能监控指标上运行指标特征提取[12]和相似性分析[13-14]算法，基于指标之间的相似性推断设备、组件之间的关联性；

(3)合并告警和指标中发现的关联关系，共同形成最终模型。

### 2.2 在告警数据集上的多层次关联性扫描

以图 3 所示的 5 条告警数据为例，存在以下问题：

(1)如果采用单次时间切片，A1 和 A2 两个告警的第一次成对出现会被切到两个时间片中，从而变成两个不相关的告警，从而降低了结果的置信度。

(2)告警是有生命周期的，当两个故障具有相关性时，除了关注它们同时发生，还要关注是否同时清除。如A3 和 A4，因为没有同时清除，可以据此降低其关联的置信度。

(3)对于频繁发生的告警，会更多地关联到其他告警，但这些关联关系中有些是无效的[15]。基础扫描为了提高

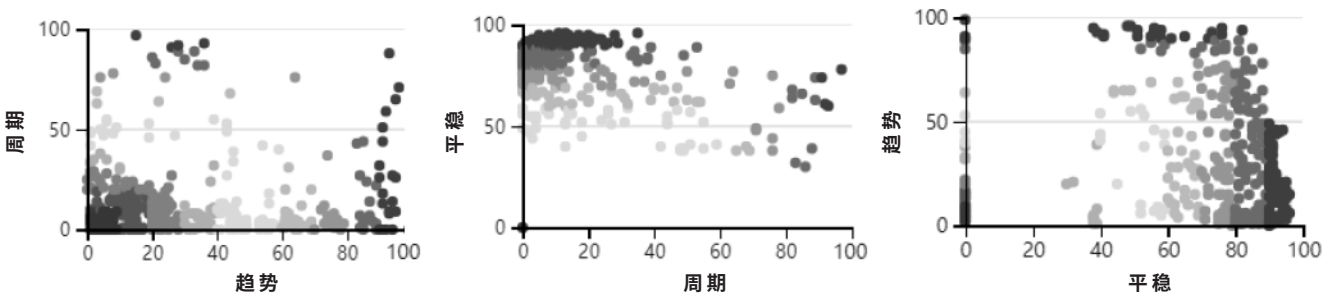



图 2 指标特征识别结果-特征分布

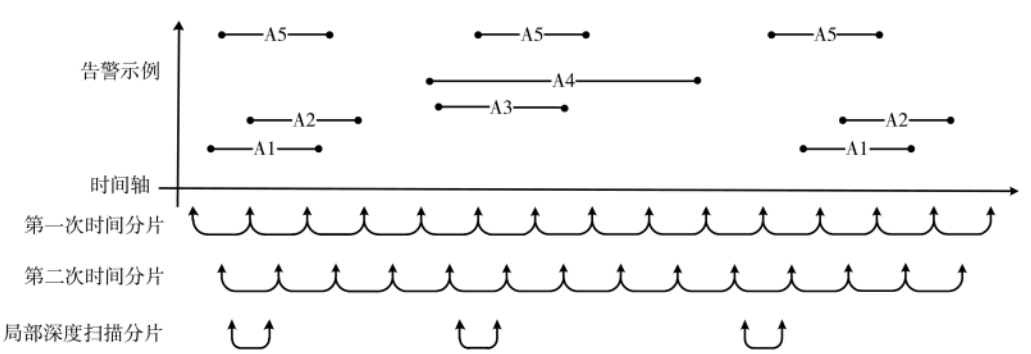



图 3 告警数据示例

敏感性，参数设置得比较宽泛，使得无效关系被纳入扫描结果，如 A5 和 A3，实际上它们只是偶然碰到一起了。

针对以上 3 个问题，本文在告警数据集上设计了 5 次扫描，前 4 次为基础扫描，如表 2 所示。第 5 次为定点深度扫描，其算法流程如图 4 所示。

表 2　基础扫描过程

| 次数 | 描述 |
|---|---|
| 第 1 次 | 从时间 $T_0$ 开始，进行周期为 $W$ 的固定时间切片，时间切片窗口和扫描参数均比较宽泛。在每个时间切片内，基于告警的发生时间收集每个时间片涵盖的告警项，作为算法输入 |
| 第 2 次 | 从时间 $T_0-W/2$ 开始，进行周期为 $W$ 的固定时间切片，其他逻辑与第一次扫描相同<br>此步解决问题(1) |
| 第 3 次 | 从时间 $T_0$ 开始，进行周期为 $W$ 的固定时间切片，时间切片窗口和扫描参数均比较宽泛。在每个时间切片内，基于告警的清除时间收集每个时间片涵盖的告警项，作为算法输入<br>此步解决问题(2) |
| 第 4 次 | 从时间 $T_0-W/2$ 开始，进行周期为 $W$ 的固定时间切片，其他逻辑与第 3 次扫描相同<br>此步解决问题(2)和问题(1) |

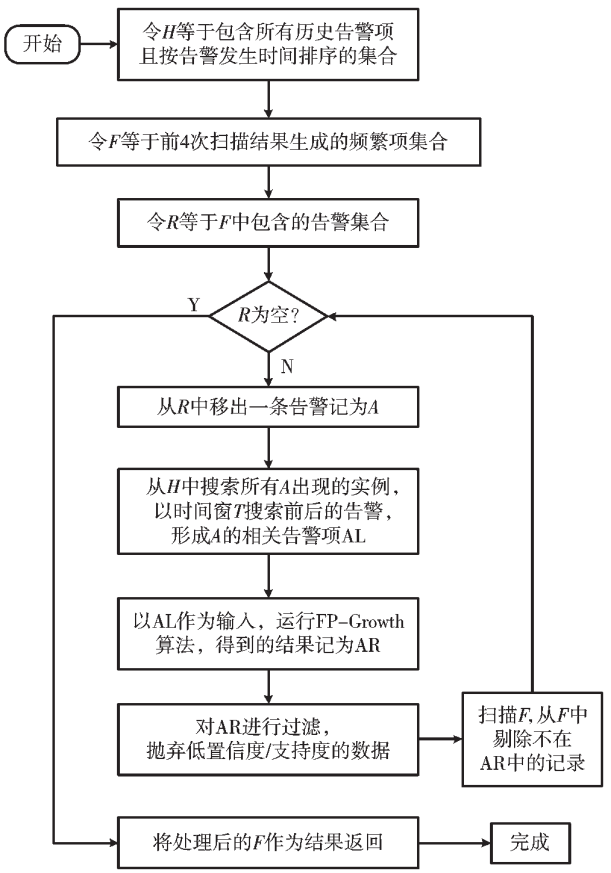



图 4　定点深度扫描算法

在完成 4 次基础扫描后，需要对扫描结果进行一个初步的合并。假设 4 次扫描结果所得关联关系集合为 $R_i$ $(i=1,2,3,4)$，则基础扫描的合并结果记为 $R_{base}$：

$$R_{base}=\mathrm{Min}(\mathrm{Max}(R_1\cup R_1)\cap\mathrm{Max}(R_3\cup R_4)) \quad (1)$$

然后，在 $R_{base}$ 上应用筛选算法 $f_1$ 来删除置信度不符合要求的关系，应用筛选算法 $f_2$ 来删除支持度不符合要求的关系，得到过滤后的结果集，记为 $R_{fBase}$。

$$R_{fBase}=R_{base}-(f_1(R_{base},\mathrm{thres})\cup f_2(R_{base},\mathrm{thres})) \quad (2)$$

其中，thres 是通过界面配置的算法参数值。

### 2.3 在性能监控指标数据集上的关联性扫描

具体到在性能监控指标数据集上的计算，其思路基于如下事实：告警是不完备数据集，因为不是所有的告警都一定发生过；监控指标是相对完备的数据集，如图5所示。

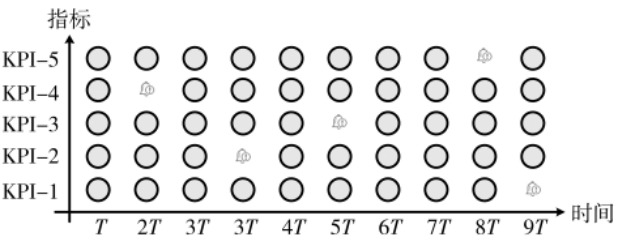



图 5　指标数据集图

图 5 中，$T$ 为指标采集周期。正常情况下，所有指标在每个采集周期均有数据，但只有个别点位会发生告警，因此指标数据比告警数据的完备度高。

从文献[16]可知，当两个设备、组件具有业务上的相关性时，一定可以表现为两个相关对象上某些监控指标的联动，包括同向上升、下降，逆向上升、下降，或者同周期波动。

例如，用户通过 Web 界面来向系统发出请求，当用户的请求数变化时，能观察到：tomcat 进程所占 CPU、数据库查询量、两台主机的负载、两个网络接口链路上的网络流量和用户的请求数有着趋势相同的变化。

根据以上分析可知，在性能监控指标数据集上的关系挖掘，就是通过比对所有指标在过去一段时间里的趋势相关性，确定指标之间的相关的置信度，从而推断设备、组件之间的关系，与告警发现的关系共同形成告警压制、溯源所依赖的关系网。

同时，本文引入了“主成分分析(PCA)”和“曲线特征提取”两步，来兼顾计算量和计算精度。

对于曲线的特征提取，本文采取斜率突变法。通过寻找合适的斜率突变点，由连接这些突变点的直线来近似拟合原始曲线，从而在保持原始曲线整体趋势特征的前提下，降低局部数据噪声。

针对每个设备、组件所获得的指标特征数据，需要进一步通过主成分分析法来进行数据降维。将一个设备、组件上的 $N$ 个指标看作是一个 $N$ 维的数据集，通过主成分分析后，筛选出 $1-M$ 个具有代表性的指标参与最终的运算。

在分别完成告警数据集和性能指标数据集上的关

联关系计算后，将两组数据进行合并，形成最终的结果。

2.4 应用效果

选择某数据中心作为试点，实现“智能化告警关联和溯源”场景研发及落地验证。对基于智能关联的告警溯源评价标准为：

(1)压缩比，指在单位时间内工具汇报的根源告警数/参与聚合的原始告警数；

(2)精确率，指在单位时间内对应实际故障的根源告警数/工具汇报的根源告警数；

(3)召回率，指在单位时间内对应实际故障的根源告警数/故障总数。

目前本文使用的测试集包含原始告警 8 757 条，经关联分析后，对其中 4 590 条告警进行了压制，推荐了 69 个根源告警/风暴，告警压缩比例为 55.7%，告警根源分析准确率约 50%，召回率约 60%，有效提高了告警的精度和有效性。

## 3 结论

在云计算和大数据快速发展的背景下，本文研究基于机器学习的智能化运维工具，将大数据技术、机器学习技术应用于中国移动一级 IT 云的运营运维工作中，可以通过机器学习的方法掌握运维数据之中的规律，自动生成更准确的阈值或通过异常模式的识别判断异常的发生，从而以机器决策分析代替传统的人工经验决策；通过处理和分析海量的运维数据、运维大数据的应用，企业能够提前发现 IT 系统中潜在的问题和风险，将被动响应式的风险处理方式变为自动性防御；通过机器学习的方式，在异常监测、告警关联压制、容量预测等环节发挥效用，提高运维的效率和质量。

根据智能运维管理的发展应用和 IT 云的运维管理需求，后续的应用重点为：探索基于智能预测的主动运维，基于模型自动预测、预警，实现对系统故障的提前感知，并可以将预警与自动处理机制对接，实现运维信息立体交换，让运维管理员获得充分的运维关联信息，从而对潜在故障进行恢复或优化；此外，启动大数据挖掘研究，不仅只针对运维数据进行分析，持续优化完善业务数据整合和动态关系建模，将现在分散在各个系统中的运维信息进行有效的整合与利用。

## 参考文献

[1] 周永道，王会琦，吕王勇.时间序列分析及应用[M].北京：高等教育出版社，2015.

[2] BREUNIG M M，KRIEGEL H P，NG R T，et al.LOF：identifying density-based local outliers[C]//Acm Sigmod International Conference on Management of Data.ACM，2000.

[3] PAPA J P，FALCÃO A X.Efficient supervised optimum-path forest classification for large datasets[J].Pattern Recognition，2012，45(1)：512-520.

[4] 高焕臣，王俊成.皮尔逊曲线拟合的完全程序化[J].海洋通报，1994(3)：62-70.

[5] 林冠群，吴裕益.相关系数与信度系数的关联及问题[J].测验学刊，2005，52(2)：29-60.

[6] 薛允莲.时间序列中移动假日效应的识别及处理[D].广州：中山大学，2009.

[7] LUO J，OUBONG G.A comparison of SIFT，PCA-SIFT and SURF[J].International Journal of Image Processing，2009，3(4)：1-10.

[8] 李靖华，郭耀煌.主成分分析用于多指标评价的方法研究——主成分评价[J].管理工程学报，2002，16(1)：39-43.

[9] 吴晓知.网络故障诊断专家系统知识库的设计与实现[D].成都：电子科技大学，2007.

[10] 于漫，胡明，金刚，等.关联规则算法的电信网络告警应用[J].吉林大学学报(信息科学版)，2010(3)：49-54.

[11] 张素琪.案例推理关键技术研究及其在电信告警和故障诊断中的应用[D].天津：天津大学，2014.

[12] PAPADIMITRIOU S.Streaming pattern discovery in multiple time-series[C]//Proc.Intl.Conf.Very Large Data Bases，2005：697-708.

[13] 刘宝生，闫莉萍，周东华.几种经典相似性度量的比较研究[J].计算机应用研究，2006，23(11)：1-3.

[14] 肖辉，胡运发.基于分段时间弯曲距离的时间序列挖掘[J].计算机研究与发展，2005，42(1)：72-78.

[15] 肖波，徐前方，蔺志青，等.可信关联规则及其基于极大团的挖掘算法[J].软件学报，2008，19(10)：2597-2610.

[16] 李开宇.一种基于时序相似度对告警信息分析查询的装置及方法：中国，CN109753526A[P].2019-05-14.

(收稿日期：2021-03-22)

作者简介：

刘虹(1972-)，女，高级工程师，主要研究方向：IT 系统设计、IT 架构、IT 应用规划。

滕滨(1976-)，男，高级工程师，主要研究方向：IT 架构、云计算。

张琳(1981-)，女，高级工程师，主要研究方向：IT 架构、云计算。

扫码下载电子文档

# 版权声明