# 大数据场景下基于机器学习的5G云网告警关联分析

## 5G Alarm Association Analysis Based on Machine Learning in Big Data Scenarios

常　铮[1]，马少伟[1]，毛斌宏[2]（1. 中国电信集团有限公司，北京 100032；2. 中国电信股份有限公司研究院，广东 广州 510630）
Chang Zheng[1]，Ma Shaowei[1]，Mao Binhong[2]（1. China Telecom Group Co.，Ltd.，Beijing 100032，China；2. Research Institute of China Telecom Co.，Ltd.，Guangzhou 510630，China）

**摘　要：**
提出一种基于关联规则挖掘算法的5G云网告警分析方案，对机器学习算法FP-Growth进行契合5G云网告警场景的改进和应用，利用现网告警数据展开告警关联分析工作，挖掘网元及云侧告警之间的关联关系，并进行告警压缩和收敛，为5G云网故障分析和定位提供有效帮助。此外，基于实际告警数据对Apriori和FP-Growth算法的性能进行了比较，结果表明，FP-Growth关联规则挖掘算法与Apriori相比效率更高，更适合海量数据场景下的告警关联分析。

**关键词：**
5G核心网；关联分析；机器学习；FP-Growth；Apriori
**doi：**10.12045/j.issn.1007-3043.2022.06.013
**文章编号：**1007-3043（2022）06-0071-06
**中图分类号：**TN919
**文献标识码：**A
**开放科学（资源服务）标识码（OSID）：**

**Abstract：**
It proposes a 5G alarm association analysis scheme in which the machine learning algorithm FP-Growth is improved and applied to fit the 5G network alarm scenario. Association analysis is carried out by using the real 5G network alarm data to mine association rules in network function and cloud infrastructure alarms. Such alarm association rules help in network fault analysis and location as well as alarm compression. In addition，it evaluates the performance of the two algorithms Apriori and FP-Growth on the basis of real 5G network alarm data. The results show that the FP-Growth algorithm is more efficient than Apriori，and is more suitable for alarm association analysis in massive data scenarios.

**Keywords：**
5G core network；Association analysis；Machine learning；FP-Growth；Apriori

**引用格式：**常铮，马少伟，毛斌宏. 大数据场景下基于机器学习的5G云网告警关联分析[J]. 邮电设计技术，2022（6）：71-76.

## 1 概述

与4G相比，5G核心网架构更加零散化、基础设施云化、网元功能虚拟化等，这使得各个网络功能能够独立地扩容和演进，并能方便地按需部署。这样的网络架构更加灵活和高效，可以满足不同业务场景的需求，但是5G核心网的网元种类数增多（见图1），每种网元有多种微服务，5G核心网CT云机房内有ToR交换机、EoR交换机、DCGW路由器和物理服务器等多种类型的设备，承载网元功能的又有多种虚机和主机，造成5GC云网告警数目剧增且种类繁多，故障分析场景复杂，定位困难。如何在海量告警之中进行关联性分析和故障定位，缩短根因定位时间，提高故障处置的及时性和有效性，同时进行告警压缩，减少派单，降低人力成本，成为亟待解决的问题。

在上述场景下，关联分析成为帮助解决告警关联和故障原因定位问题的有效手段[1-5]。关联规则挖掘能够从数据集中发现项与项之间的关联关系，并从中提取出符合一定条件的强关联规则进而指导业务。典型算法有Apriori算法[6-7]和FP-Growth算法[8-9]，

收稿日期：2022-04-14

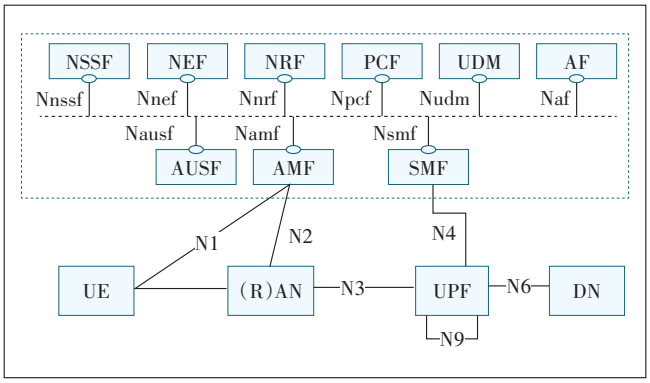


**图1**　非漫游场景的5G网络架构（服务化接口形式）

Apriori算法基于广度优先，需要多次扫描数据集获取频繁项集，且会产生大量的候选集，算法效率较低。FP-Growth算法基于深度优先，通过压缩数据集构造频繁模式树FP-Tree，通过提取树的频繁项集可以有效减少对数据库的扫描，极大提高运算效率。本文提出基于关联规则挖掘算法的5G云网告警分析方案，对FP-Growth算法进行契合5G云网告警场景的改进和应用，通过实例分析介绍FP-Growth算法挖掘5G云网告警关联规则的原理以及结果，同时利用现网告警数据将FP-Growth与Apriori算法的性能及效率进行对比。该方案对5G云网告警进行清洗整合，通过关联规则挖掘算法分析不同告警之间的关联关系，得到告警关联规则，结合专家对关联规则的解释、筛选和打标，完成5G云网告警关联规则库的构建，将该告警关联规则库用于日常派单时的告警压缩及故障发生时的告警分析和问题定位，从而降低人力分析成本，提高事件处置效率，提升网络运行维护智能化水平。

## 2 关联规则挖掘算法原理及概念

### 2.1 关联分析基本概念

#### 2.1.1 项集（Item Set）

数据库中的最小单位信息为项，项集即项的集合。对于5G云网告警分析来说，就是若干个5G云网告警的集合，包含$k$个项的项集称为$k$-项集。

#### 2.1.2 事务数据库（Transaction Database）

假设$I=\{i_1,i_2,\cdots,i_q\}$为事务数据库中所有项的集合，事务$T$是项集$I$中项的集合，满足$T\subseteq I$且非空。事务数据库$D$为所有事务的集合。对于5G云网告警分析来说，$I$就是一次分析中所有告警类型的集合，事务$T$就是根据时间关系或资源关系划分的一个告警组，事务数据库$D$就是这样所有告警组的集合。

#### 2.1.3 关联规则（Association Rules）

关联规则是形如$X\Rightarrow Y$的蕴含式，其中$X=\{x_1,x_2,\cdots,x_m\}\subset I$，$Y=\{y_1,y_2,\cdots,y_n\}\subset I$，$X$和$Y$是不同的项集且非空。$X\Rightarrow Y$表示在一次事务中如果前件$X$出现，那么后件$Y$会有一定概率出现。例如，对5G云网告警关联分析来讲，关联规则$X\Rightarrow Y$表示集合$Y$中的告警以一定概率随着$X$集合中告警的出现而出现。

#### 2.1.4 关联强度指标

支持度（Support）：假设有关联规则$X\Rightarrow Y$，支持度Support$(X\Rightarrow Y)$表示某项集$X\cup Y$在事务数据库中出现的可能性：

$$\text{Support}(X\Rightarrow Y)=P(X,Y) \tag{1}$$

置信度（Confidence）：假设有关联规则$X\Rightarrow Y$，置信度Confidence$(X\Rightarrow Y)$表示在前件发生的条件下后件发生的概率，是一个条件概率：

$$\text{Confidence}(X\Rightarrow Y)=P(Y|X)=P(X,Y)/P(X) \tag{2}$$

#### 2.1.5 强关联规则

强关联规则是关联强度指标满足一定关联强度条件的规则，一般定义强关联规则为满足最小支持度且满足最小置信度阈值的规则。

### 2.2 Apriori及FP-Growth算法原理

Apriori算法伪代码如表1所示，利用$k$-项集搜索$(k+1)$-项集，频繁项集自连接生成候选集，候选集再剪枝生成频繁项集，一直到没有新的频繁项集产生为止。从伪代码中可以看到算法的执行过程中，Apriori算法需要对事务数据库反复扫描，且会产生大量的候选项集，效率较低。

**表1**　Apriori算法伪代码

| |
|---|
| 输入：事务数据库$D$，最小支持度MinSupport<br>输出：频繁项集FreqItemSet |
| 算法过程：<br>①令$k=1$<br>②扫描数据库，计算支持度得到频繁$k$-项集FreqItemSet_$k$<br>③如果频繁$k$-项集为空，结束，输出所有频繁项集FreqItemSet；如果不为空，转到④<br>④由频繁$k$-项集生成$k+1$阶候选集<br>⑤对$k+1$阶候选集进行剪枝<br>⑥令$k=k+1$，转到② |

FP-Growth算法通过构造频繁模式树这种比较紧凑的数据结构，将频繁模式信息进行压缩，本质上是一种深度优先搜索算法[10]，只需对事务数据库进行2次扫描就可以获取频繁模式集合。FP-Growth算法主要分为构建频繁模式树和挖掘树中的频繁模式2个过程，其算法伪代码如表2所示。本文第4章的实例分

表2　FP-Growth算法伪代码

| |
|---|
| 输入：事务数据库 $D$，最小支持度 MinSupport<br>输出：频繁项集 FreqItems |
| 算法过程：<br>①第一次扫描事务数据库，剔除不满足支持度的项，得到频繁1-项集并按支持度降序排列，创建项头表[11] headerTable<br>②创建根节点 $T$（null），第2次扫描事务数据库，调用 updateTree 函数生成 FP-Tree 并更新项头表<br>③令 FreqItems 为空列表，preFix 为空集<br>④调用 mineTree(headerTable，prefix，FreqItems) 函数，进行频繁模式挖掘[12]，对于项表头中每一个频繁项 $\alpha_i$：<br>•令 newFreqSet = preFix，newFreqSet = newFreqSet ∪ $\alpha_i$，FreqItems 列表追加 newFreqSet<br>•获取 $\alpha_i$ 对应的条件模式基 conditionalPattBase，利用条件模式基构建新的频繁模式树 conditionalTree，并获得项头表 newHeaderTable<br>•如果 newHeaderTable 为空，输出 FreqItems，结束；如果不为空，递归调用 mineTree(newHeaderTable，newFreqSet，FreqItems) |

析中，将利用现网5G云网的告警数据，测试比较Apriori算法和FP-Growth算法的运行效率。

## 3 基于关联规则挖掘算法的5G云网告警分析

随着现网规模的扩大，5G的网元、设备等每时每刻都在产生大量的告警。当某个网元或设备发生异常时，往往与故障相关联的多个网元、设备以至业务流程也会产生相应的告警信息，短时间内大量的告警事件相互叠加，经常会将最关键的告警信息淹没，造使故障分析和根源定位困难。本文提出基于关联规则挖掘算法的5G云网告警分析方案，在海量告警中发掘有效信息以提高生产效率。如图2所示，方案包括了生产系统、5G云网告警实时监测系统、存储系统、关联规则挖掘系统、规则匹配部分以及规则利用部分。

a）生产系统。它是指全国每时每刻都在运行的5G现网，包括了无线侧、核心网以及承载网的部分。生产系统每时每刻都在源源不断地产生大量的5G云网告警。

b）5G云网告警实时监测系统。它是指通过网元、设备或者网管进行实时告警监测、采集和处理的过程。通常采集服务通过SNMP、TCP、SSH以及SSL等协议连接网元，通过主动监听或被动接收的方式采集告警，并将不同厂家的告警进行格式统一等处理。本文重点分析包括5G核心网网元以及5G核心网机房内数据通信设备（交换机、路由器等）、物理主机及虚拟主机等设备产生的告警，即5G网络侧和云侧告警。

c）存储系统。存储系统中存储了包括网元及设备告警、日志、性能、配置信息和资源拓扑等。从生产系统产生的5G云网告警通过实时监测系统采集处理送入存储系统分类存储。本文的方案选择Elastic-Search集群作为5G云网告警的存储系统，5G网侧和云侧的告警被分类存储在不同的索引下。

d）关联规则挖掘系统。主要具有如下功能。

（a）告警数据预处理模块从存储系统不同的数据库表中提取5G云网告警，根据业务层次将不同类型告警进行对齐和归一化处理，并进行清洗、聚合及排序等。

（b）训练相关数据转换模块将预处理模块生成的告警根据时间或资源关系进行合并处理，将原始告警转换为5G云网告警事务数据库，以便之后的关联规则挖掘算法进行模型训练。

（c）算法模型的运行主要分为2个部分：第1个部分是关联规则挖掘算法（Apriori、FP-Growth等）从5G云网告警事务数据库中挖掘出频繁项集的过程，第2个部分是根据业务需求从频繁项集中形成强关联规

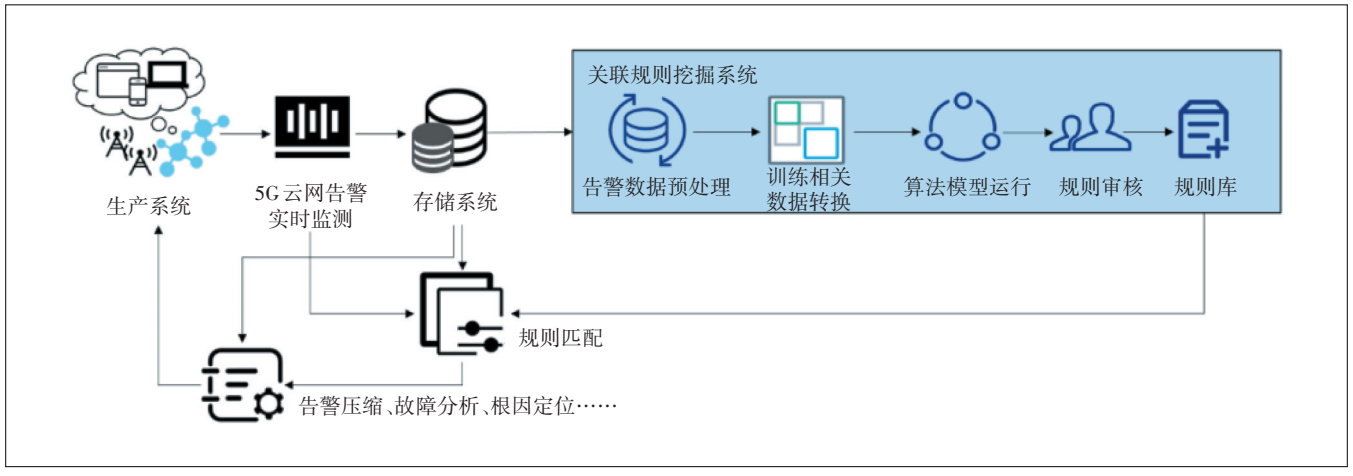


图2　基于关联规则挖掘算法的5G云网告警分析方案

则的过程。

（d）规则审核模块是云网专家对算法模型运行生成的强关联规则进行审核和打标的过程。

（e）规则存储模块将专家审核通过的强关联规则进行编码入库，以便后续进行规则匹配和规则利用。

e）规则匹配模块。根据实时监听的5G云网告警、告警规则库中的5G云网告警关联规则以及从存储系统中获得的网元、设备及虚机主机等之间的资源拓扑关系进行规则匹配分析，生成匹配结果。

f）规则利用模块。根据规则匹配结果进行告警压缩作用于生产系统，并且利用存储系统中的资源拓扑关系、性能指标等对规则匹配结果进行再次分析和加工，用于故障分析和根因定位，从而提升维护效率，降低人力和时间成本。

## 4 5G云网告警关联分析实例

### 4.1 分析流程

本节重点研究如何在数据预处理、FP-Growth算法强关联规则生成以及算法实现方式等方面进行适合5G云网告警场景的适配和改进，从而更有效地解决告警关联分析问题。本节实例采用某运营商5G现网环境某设备厂家全国所辖省份2周的5G云网告警数据进行关联分析，主要分为以下几个过程。

a）告警数据预处理阶段解决源数据存在的冗余、不完整或不统一等问题，提取5G云网告警之后首先进行数据清洗，对一些主要字段（如告警编码、告警标题、告警设备等）有缺失的告警数据进行剔除，其次进行数据集成，将不同索引来源的告警进行整合，根据业务层次和意义进行字段对齐等归一化处理，对一些意义不大的低层级通知告警进行过滤。

b）数据转换部分根据告警发生的时间关系将预处理后的告警进行项集转换，形成5G云网告警事务数据库。相比于很多关联规则挖掘应用中使用简单的时间窗口切分法，本文采用滑动时间窗口方式进行5G云网告警事务数据库的生成。滑动时间窗口示意图如图3所示，划定一个时间窗口长度并指定一个滑动窗口步长，就可以得到多个时间窗，每个时间窗内对应的告警都可以经过处理成为告警事务数据库中的项集，这种方式可以有效防止有关联关系的告警被划分到不同的项集中。

c）关联规则挖掘部分包括挖掘频繁项集和生成强关联规则，关联规则挖掘算法FP-Growth根据最小支持度挖掘符合支持度阈值的频繁项集，关联规则生成部分根据设置的强关联规则条件从频繁项集中筛选出强关联规则。传统的FP-Growth算法在应用时通常采用支持度和置信度作为关联强度的衡量指标，但仅用这2个指标不足以满足5GC云网告警的关联分析需求，甚至有时仅从支持度和置信度得到的强关联规则$X \Rightarrow Y$中，$X$对$Y$具有抑制作用[13]。本文在进行强关联规则的挖掘和提取时增加了提升度、杠杆率和确信度3种关联强度指标，提升了挖掘出的告警关联规则的准确性和有效性。

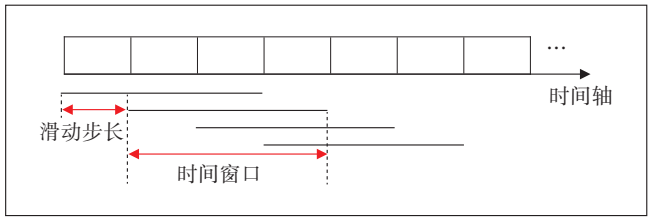


图3　滑动窗口示意图

（a）提升度（Lift）。提升度Lift($X \Rightarrow Y$)用来衡量当$X$出现的情况下，对$Y$出现的概率是否有提升。当提升度为1时，$X$和$Y$相互独立；当提升度大于1时，$X$和$Y$正相关，越大正相关性越强；当提升度小于1时，$X$和$Y$负相关，越小负相关性越强[14]。

$$\text{Lift}(X \Rightarrow Y) = P(Y|X)/P(Y) \quad (3)$$

（b）杠杆率（Leverage）。杠杆率与提升度类似，都是衡量在前件出现的条件下，是否对后件出现的概率有提升。杠杆率Leverage($X \Rightarrow Y$)为0时，$X$和$Y$相互独立；杠杆率大于0时，$X$和$Y$正相关，越大正相关性越强；杠杆率小于0时，$X$和$Y$负相关，越大负相关性越强。

$$\text{Leverage}(X \Rightarrow Y) = P(Y|X) - P(Y) \quad (4)$$

（c）确信度（Conviction）。当确信度Conviction($X \Rightarrow Y$)为1时，$X$和$Y$相互独立；当确信度大于1时，$X$和$Y$正相关，越大正相关性越强；当确信度小于1时，$X$和$Y$负相关，越小负相关性越强。

$$\text{Conviction}(X \Rightarrow Y) = \frac{1 - P(Y)}{1 - \text{Confidence}(X \Rightarrow Y)} \quad (5)$$

d）调参部分主要是针对c）中的关联规则挖掘结果调整最小支持度和最小置信度的过程。在挖掘的过程中根据告警数据特点和经验设置初始的最小支持度和最小置信度，分析挖掘后生成的5GC云网告警关联规则是否合理，根据关联规则是否过少或者是否冗余的情况再进行参数的调整并重新进行挖掘，直至

生成合理且满足业务需求的关联规则挖掘结果。

此外，通过调研发现很多FP-Growth算法的实现方法是参照参考文献[15]中的思路，采用frozenset的方式并根据频次进行排序。frozenset是一种Python数据结构，frozenset中元素的存储顺序存在不确定的情况，且仅用频次作为排序依据容易引起排序不稳定问题，进而影响频繁项集和关联规则的结果，这对于海量云网告警的关联分析是极其不利的。因此本文采用有序列表方式记录项及其频次，并在排序时采用频次和项值双重关键字的形式，避免不稳定排序，保持FPTree结构的稳定一致，从而保证频繁项集和关联规则结果的稳定和准确。

### 4.2 分析结果

关联规则挖掘算法从频繁多项集中生成的众多强关联规则需经过专家审核打标后才能入库存储，形成运维经验，从而进行现网应用，本文经过专家审核通过的部分关联规则的样例如表3所示。

**表3**　专家审核通过的部分关联规则（样例）

| 序号 | 前因 | 后果 | 提升度 | 杠杆率 | 置信度 | 确信度 |
|---|---|---|---|---|---|---|
| 1 | SCP-http对等端故障 | SCP-http对等端组故障 | 6.021 | 0.048 | 0.354 | 1.456 |
| 2 | UPF-网卡端口down检测告警 | UPF-sriov网卡速率小于10gbps告警 | 8.172 | 0.098 | 0.950 | 17.720 |
| 3 | AMF-udm不可达 | AMF-ausf不可达 | 482.000 | 0.002 | 1.000 | inf |
| 4 | switch-接口二层协议态down | switch-接口IPv4协议态down | 51.233 | 0.016 | 0.833 | 5.902 |

规则1，SCP网元的“http对等端故障”与“http对等端组故障”具有强关联关系，这是因为当网元的所有http对等端都发生故障后，会导致“http对等端组故障”告警产生。

规则2，UPF网元的“sriov网卡速率小于10gbps告警”与“网卡端口down检测告警”具有强关联关系，而且置信度较高，达到了0.959，实际上当UPF网卡发生故障时，“网卡端口down检测告警”与“sriov网卡速率小于10gbps告警”往往一同产生。

规则3，AMF网元的“udm不可达”告警与“ausf不可达”告警具有强关联关系，现网存在UDM和AUSF合设的情况，因此当UDM发生异常不可达，AMF也会产生到AUSF不可达的告警。

规则4，交换机设备的“接口的二层协议down”告警和“接口IPv4协议态down”告警具有强关联关系，二层协议down会引起接口的IPv4协议down。

经专家审核通过的强关联规则往往存在协议原理上、拓扑结构上以及业务流程上的关联，将这些沉淀下来的规则反作用于现网，可以用于告警压缩、根因定位以及故障分析，提高网络维护效率，降低运维的人工依赖性，提升网络运营的智能化水平。

本文也利用现网告警数据对Apriori和FP-Growth 2种关联规则挖掘算法的实际运行速度进行了对比，2种算法的效率对比如图4所示。因为关联规则挖掘算法重点在于挖掘频繁项集，频繁项集相同，强关联规则条件相同，生成的规则就相同，因此这里对比的是Apriori和FP-Growth 2种算法从事务数据库中挖掘出频繁项集所用的时间。从图4可以看出，2种算法的运行时间随着数据量的增大而增大，随着支持度的增大而减小，因为支持度越大，算法实际处理的数据量会越少，因此算法的挖掘速度就会越快。Apriori算法的运行时间随数据量的增加而增长的速度比FP-Growth快很多，也即Apriori算法的性能受数据量的影响较大，数据规模越大，Apriori算法的运行效率越低，这是因为Apriori算法需要多次扫描数据库，而相比于Apriori，FP-Growth算法的运行时间随数据量的增加一直处于较低水平，性能受数据量的影响较小，效率更高。在进行大数据规模的海量告警关联规则分析时，FP-Growth比Apriori算法更为合适，因此本文基于FP-Growth算法探索适合于5G云网告警场景的关联分析和应用。

## 5 结束语

当今5G业务快速发展，运营商现网运营中5G云网告警数目剧增且种类繁多，故障分析场景复杂，本文提出了一种基于关联规则挖掘算法的5G云网告警分析方案，对FP-Growth算法进行契合5G云网告警场景的改进和应用，并利用现网的告警数据进行了实例分析和规则展示。方案通过关联规则挖掘算法，在海量的云网告警中分析告警关联性，挖掘有价值的强关联规则进行专家经验沉淀，用于告警压缩、根因定位等网络运营维护及故障诊断场景。此外，本文还利用实际的5G云网告警数据对Apriori和FP-Growth 2种关联规则挖掘算法的运行效率进行了对比，实验表明FP-Growth算法性能相比Apriori受数据量的影响更小，挖掘效率更高。通过部署基于关联规则挖掘算法的5G云网告警分析方案，可以实现运维经验的迭代，

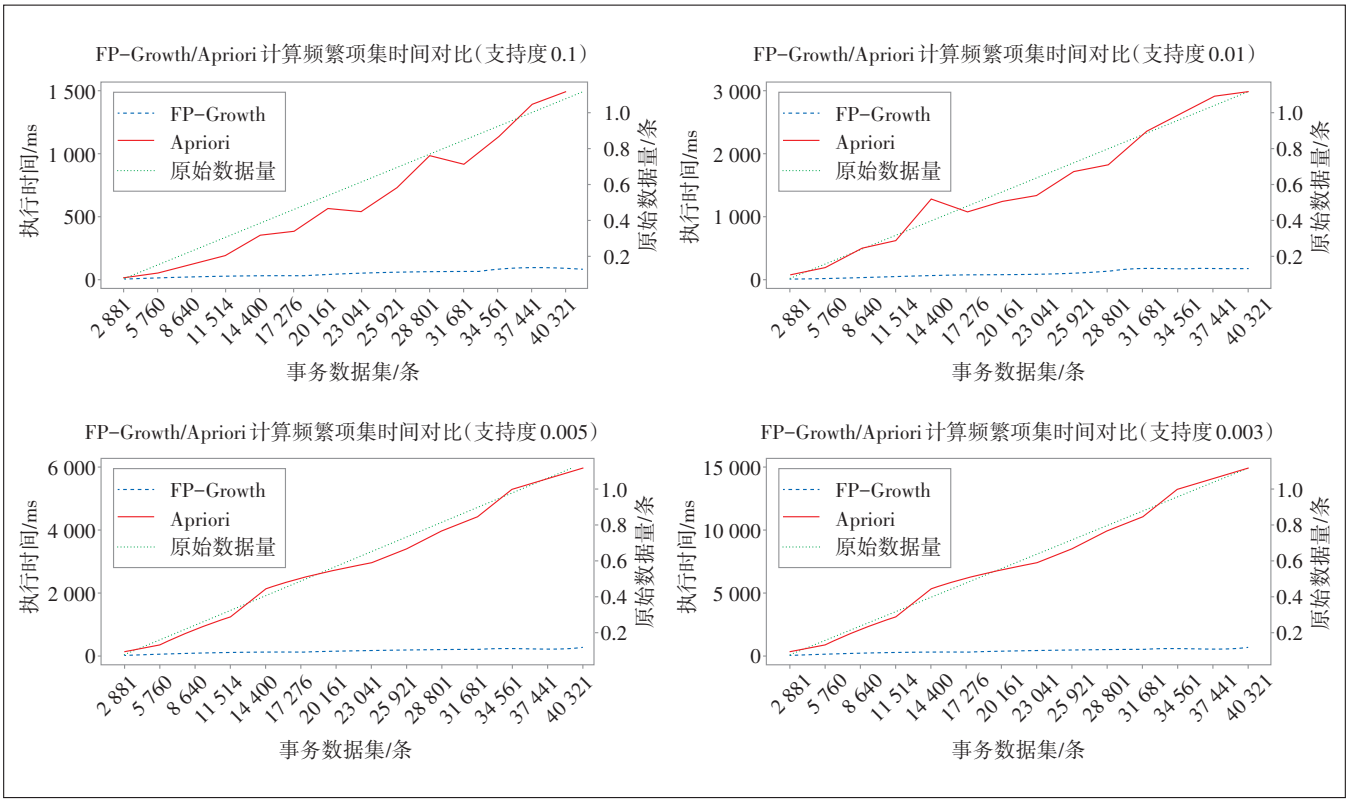


**图4**　不同数据量、不同参数下 Apriori 和 FP-Growth 算法效率对比

降低人工依赖性，为5G云网的数字化运营注智赋能。

## 参考文献：

[1] 柳媛. 基于无监督学习的故障分析研究[D]. 北京：北京邮电大学，2021.

[2] 李川. 基于Spark的电信网络告警大数据关联规则算法研究与实现[D]. 北京：北京邮电大学，2017.

[3] 张雄. 基于关联规则的电信网告警相关性分析[D]. 南京：东南大学，2016.

[4] 张永华. 基于大数据技术的电信网络告警关联分析设计与实现[J]. 电信工程技术与标准化，2016，29(4)：18-23.

[5] 杨磊. 基于FPgrowth机器学习的影响用户感知无线根因问题的快速定位方法研究[J]. 江苏通信，2019，35(2)：56-62.

[6] AGRAWAL R，IMIELIŃSKI T，SWAMI A. Mining association rules between sets of items in large databases[J]. ACM Sigmod Record，1993，22(2)：207-216.

[7] AL-DHARHANI G S，OTHMAN Z A，BAKAR A A. A graph-based ant colony optimization for association rule mining[J]. Arabian Journal for Science and Engineering，2014，39(6)：4651-4665.

[8] SAHOO J，DAS A K，GOSWAMI A. An effective association rule mining scheme using a new generic basis[J]. Knowledge and Information Systems，2015，43(1)：127-156.

[9] JIAO M H，YAN P，JIANG H Y. Research and application on Web information retrieval based on improved FP-growth algorithm[J]. Wuhan University Journal of Natural Sciences，2006，11(5)：1065-1068.

[10] YUAN Y，HUANG T. A matrix algorithm for mining association rules[C]//International Conference on Intelligent Computing. Springer，Berlin，Heidelberg，2005：370-379.

[11] JAMSHEELA O，RAJU G. An adaptive method for mining frequent itemsets efficiently：An improved header tree method[C]//2015 International Conference on Advances in Computing，Communications and Informatics（ICACCI）. Kochi，India：IEEE，2015：1078-1084.

[12] YUAN J B，DING S L. Research and improvement on association rule algorithm based on FP-growth[M]//WANG F L，LEI J，GONG Z，et al. Web information systems and mining. Berlin，Heidelberg：Springer Berlin Heidelberg，2012：306-313.

[13] HAN J W，KAMBER M，PEI J. Data mining：concepts and techniques[M]. Burlington，MA：Morgan Kaufmann，2012.

[14] LI Y F，LI Y S. E-commerce order batching algorithm based on association rule mining in the era of big data[C]//2018 Chinese Control and Decision Conference（CCDC）. Shenyang，China：IEEE，2018：1934-1939.

[15] HARRINGTON P. Machine learning in action[M]. Shelter Island，N. Y：Manning Publications Co，2012.

**作者简介：**

常铮，毕业于北京邮电大学，硕士，主要从事网络智能化运维工作；马少伟，毕业于北京交通大学，工程师，硕士，主要从事5G网络管理等工作；毛斌宏，毕业于哈尔滨工程大学，高级工程师，高级企业信息管理师，硕士，主要从事电信运营支撑系统研究、5G网络管理研究等工作。