DOI: 10.19783/j.cnki.pspc.200715

分布式并行 FP-growth 算法在二次设备缺陷监测中的应用

方晓洁，黄伟琼，叶东华，黄宇柏

（国网福建省电力有限公司漳州供电公司，福建 漳州 363000）

摘要：智能变电站设备监控数据存储分散，主站获取设备缺陷特征的难度大，有必要通过分布式数据挖掘的方法发现设备缺陷和信号之间的关系。分布式并行频繁模式树(FP-growth)算法采用 Hadoop 框架和 Mapreduce 算法，能够快速有效地发现信号间的强关联关系。针对二次设备的缺陷特征，建立异常模型，提出遥信历史数据准备和清洗方法，滤除复归、抖动等噪声信号，并将字符串数据转换为以关键字为标识的事务数据项集。在此基础上采用分布式并行 FP-growth 算法挖掘各变电站历史数据库异常信号的频繁项集和强关联关系。应用结果表明，该方法能够有效地发现二次设备的频发异常，找到诱发异常的缺陷，为家族性缺陷的认定提供数据基础。

关键词：分布式并行频繁模式树；数据挖掘；关联规则；频发异常；家族性缺陷

Application of a distributed parallel FP-growth algorithm in secondary device defects monitoring

FANG Xiaojie, HUANG Weiqiong, YE Donghua, HUANG Yubai

(Zhangzhou Power Supply Company, State Grid Fujian Electric Power Co., Ltd., Zhangzhou 363000, China)

Abstract: The monitoring data of smart substation devices is stored in a decentralized way. It is difficult to extract equipment defect features, so it is necessary to determine the association rules between device abnormal signals and defects by data distributed mining. Using a Hadoop framework and the MapReduce algorithm, a distributed parallel FP-growth algorithm can quickly and effectively find the strong correlation between signals. Given the defect characteristics of the secondary devices, an abnormal model is established. A method of preparing and cleaning the historical data of remote signals is proposed. This can filter out noise signals such as reset and jitter, and convert the string data into the key words data item set. The distributed parallel FP-growth algorithm is used to mine the frequent item sets and find strong correlation of abnormal signals in the historical database of each substation. The application results show that this method can effectively determine frequent abnormalities in the secondary devices and find the defects, providing the data basis for the identification of family defects.

This work is supported by the National Key Research and Development Program of China “Key Technologies Research and Standardization of IoT Terminal Evaluation Platform” (No. 2018YFB21002).

Key words: distributed parallel FP-growth; data mining; association rules; frequent abnormality; family defect

0 引言

智能变电站站端信息全面，包含了大量反映二次设备运行状态的数据。因此，多地建设了面向自动化设备运行检修的运维监控系统，运维监控系统主要采集自动化系统的运行状态信号，并应用多种算法识别缺陷，提供远方监视、诊断和管理等功能[1-6]。

变电站智能电子设备具有丰富的自我感知功能，能够为外界提供表征其健康状态的装置通信状态、自检、异常等告警类信息，以及内部温度、工作电压、光口光强等量测类数据，还包括定值、压板、装置参数等配置类信息[7-9]。这些信息与设备缺陷之间存在一定的关联关系，然而这些数据主要存储于变电站端，只有极少量的信号上送主站。二次设备运维人员几乎无法获取表征智能设备实时运行状态的重要特征量。

相对于主站，变电站端的数据更全面、更丰富多样化，有利于提取设备特征信息、挖掘智能设备缺陷和信号之间的关联规则。关联规则是指两个及

基金项目：国家重点研发计划专项资助“物联网终端评测平台关键技术研究及标准化”（2018YFB21002）

以上要素项集的显性或隐性的关联或相关性，两者之间不一定具有明显的因果关系。变电站二次设备告警或缺陷数据蕴含了大量的有价值的信息，但目前仅采用数理统计、简单的比对、因果推理等方法对其处理，无法准确挖掘出数据之间的隐性关系。文献[10]提出了一系列数据挖掘算法，在智能电网多源数据分析中得到了广泛应用[11-16]，其中包括 Apriori 和频繁模式树(Frequent Pattern Tree，FP-growth)两种关联规则挖掘算法以及其衍生算法[17-23]。近年来，随机森林作为机器学习的一种新型算法在数据挖掘中也得到了广泛应用，但其更偏重于预测，而非关联规则的挖掘[24]。

文献[19]利用配电网多源数据，筛选出断线故障诊断的主要特征，采用 FP-growth 算法挖掘分支线断线不接地故障与故障位置之间的关系，有效找出了故障点。文献[20-21]均采用 Apriori 算法挖掘二次设备缺陷，获得了二次设备缺陷的几个关键属性，进一步找出了引发缺陷的原因，甚至能够为家族性缺陷的认定提供数据支撑，但从算法效率的角度，原始的 Apriori 算法消耗的计算资源较多，因此以上文献均对算法进行了改进。文献[22]提出使用 FP-growth 算法对生产管理系统的变电站二次设备缺陷进行分析，构建了缺陷模型，发现了变电站、二次设备、缺陷原因等因素之间的关系，为设备运维和消缺工作提供了有价值的信息。文献[23]提出了一种改进 FP-growth 算法，引入分布式 Hadoop 框架挖掘大数据信息，提升运算的效率。

以上均是对变电站监控系统历史数据或生产管理系统数据进行离线的数据挖掘，数据需人工预处理，且时间上有一定的延迟，本文在应用分布式并行 FP-growth 分析功能关联库的基础上，提出一种在主子站端动态协同挖掘二次设备缺陷与状态数据之间关系的方法，分析可能存在的设备缺陷。

1　智能设备异常诊断架构及模型设计

1.1 设备异常诊断架构

变电站智能设备诊断架构由数据采集、数据存储、数据预处理、数据挖掘和缺陷诊断五部分组成，如图 1 所示。其中数据采集部分沿用自动化设备运维监控系统，由测控、保护及其他装置经站控层网络上送遥信、告警、事件等数据。数据存储于变电站本地监控系统 MySQL 数据库中。变电站同时部署分布式数据文件系统(Hadoop Distributed File System，HDFS)，HDFS 从 MySQL 数据库读取遥信历史数据表。在数据预处理阶段，首先对数据进行清洗，将告警、事件等数据转换为以关键字为标识

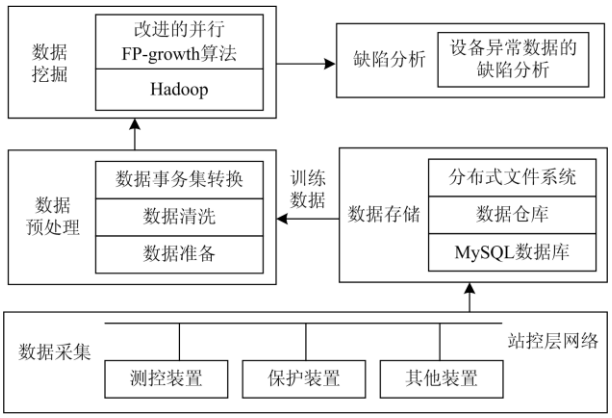

图 1 智能设备缺陷分析架构

Fig. 1 Defect analysis structure of intelligent device

的事务数据集。然后在 Hadoop 框架下，基于 Mapreduce 对各变电站节点事务数据集进行并行 FP-growth 频繁项挖掘，最后根据挖掘的结果分析设备可能存在的缺陷。

1.2 异常二次设备诊断模型

变电站实时监测数据量大、特征值多，需从中筛选出有价值的特征项用于关联规则的挖掘。系统的遥信历史数据表是主要的数据源，遥测表内的一些特征量也可以通过设定越限值等措施将异常数据转换为遥信数据。

原始遥信历史数据表中二次设备的特征项包括遥信信号名称、设备名/类型、所属变电站、发生时间、信号来源、生产厂家、信号性质等。这些遥信信息有的与电网运行状态相关，如电压越限、频率越限等；有的与通信状态有关，如 SV 断链、站控层 A 网中断；有的与设备硬件相关，如 RAM 自检出错、保护 CPU 插件异常；有的与设备的配置相关，如定值自检出错；有的与二次回路有关，如 CT 断线、对时异常等。一个异常/缺陷事件的发生会触发哪些遥信信号取决于事件间的关联度，也取决于系统/设备开发人员的设计和工程人员的配置，多带有一定程度的随机性。这些遥信信号之间可能隐藏着一定的关联关系而未被关注。

基于监控系统数据，能够通过 FP-growth 算法挖掘智能设备的频发异常、异常原因以及可能会导致异常的其他信息。从二次设备特征项中选取重要特征项进行建模，包括异常名称、设备名/类型、发生时间、生产厂家、告警性质等。针对以上特征量对智能设备的遥信量进行建模，D 是任务相关的异常数据库事务的集合，异常模型可以表示为元组形式。

$$D_i = Q(q_1, q_2, n_1, n_2, n_3) \tag{1}$$

式中：i 为整数且大于或等于 1；Q 表示特征维度集合，用于确定数据挖掘的维度，包含 5 个具体维度：q_1 表示异常的发生日期，即时间维度；q_2 表示异常发生的变电站；n_1 表示设备名称；n_2 表示设备生产厂家；n_3 表示设备异常类型。对于一条告警信息主要挖掘这 5 个维度的信息，以异常信号名称为目标进行频繁项集的搜索。

2 数据预处理

2.1 数据准备及清洗

原始的遥信历史数据表可能包含数据抖动引发的重复信号、系统自动触发的复位信号等噪声数据，因而需要对原始数据进行清洗，以获取一张反映设备状态发生次序的纯净的遥信数据表。数据预处理是异常关联规则挖掘成功的关键。

原始数据清洗的流程如图 2 所示。首先链接至数据库遥信历史数据表，采用数据库模糊查询的方法从信号的中文描述域中提取“复归”关键字，若信号为复归信号，则予以滤除。进一步读取此信号前 1 s 内的遥信数据，与此信号的中文描述进行比对，若为同一信号，则予以滤除。可以得到没有复归信号、没有抖动信号的遥信历史数据表。此时，处理后的遥信历史数据表内还有大量的中文描述，中文模糊搜索将消耗大量的计算资源。因此，还需要将遥信历史数据表进一步处理成一个只包含关键字的事务数据集。

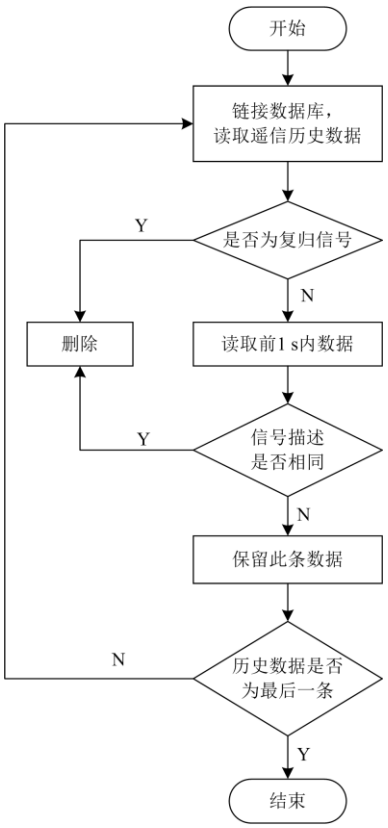

图 2 原始数据清洗方法

Fig. 2 Data cleaning method of original data

异常信号事件顺序记录(Sequence of Event, SOE)的分辨率为毫秒级，可以考虑将秒级范围内发生的异常或遥信信号组合成为一个项集，项集内的数据项可能存在关联关系，也可能不存在关联关系；所有的这些项集构成一个样本集。理论上，每一类智能设备运行状态异常的遥信信号就是一个信号项。与设备运行状态相关的信号项约有 20 多项，可以归纳总结为：EEPROM/ RAM/ FLASH 自检出错、定值自检出错、程序校验出错、插件异常、光口接收或发送功率越下限、对时异常、SV/GOOSE 链路中断、SV/GOOSE 接收不匹配、SV/GOOSE 延时越限、SV 采样丢帧、设备参数错、系统配置错、电源异常、软压板错、网卡故障、保护通道异常、CPU 温度不确定状态等。这些均为信号的中文描述，能够转换为较容易挖掘的事务数据集。

2.2 事务数据集转换

原始数据清洗后仍是离散的数据表，为便于挖掘将其转换为一系列的事件集合，即事务数据集。变电站事件、告警等遥信信号的 SOE 分辨率为 1 ms，如果若干个信号之间存在强关联关系，遥信信号相继发生的时间尺度应为毫秒级。考虑到可能存在一定的延时，则认为 1 s 内发生的信号是此时间段相继发生的信号，这些信号组成一个事务数据集。

转换的过程如图 3 所示，首先按次序逐条读取清洗后的遥信信号历史数据表，以遥信信号的 SOE 为时间标度，若此条信号前 1 s 内存在其他信号，则此条信号与其他信号组成一个事务数据集；反之，新建一个事务数据集。项集中存储的数据项以信号的中文描述关键字标识。

遍历清洗后的遥信历史数据表，直至历史数据被读取完毕。通过以上操作，完成数据的预处理，获得一个以中文描述关键字为标签的事务数据集文本，便于对遥信历史数据进行快速挖掘。

表 1 是某 220 kV 变电站监控系统某一时段内的遥信历史数据。按照上述数据预处理方法，首先清除复归、抖动信号，再筛选出与当前事件发生时间间隔不超过 1 s 的信号，提取每条信号名称的中文关键字，组成一个事务数据集。那么，表 1 中的数据可转换为 3 个事务数据集。其中 D_1={服务器运行

状态恢复}，D_2={出线控制回路异常，母线保护启动，测控装置闭锁}，D_3={母线保护录波就绪}。

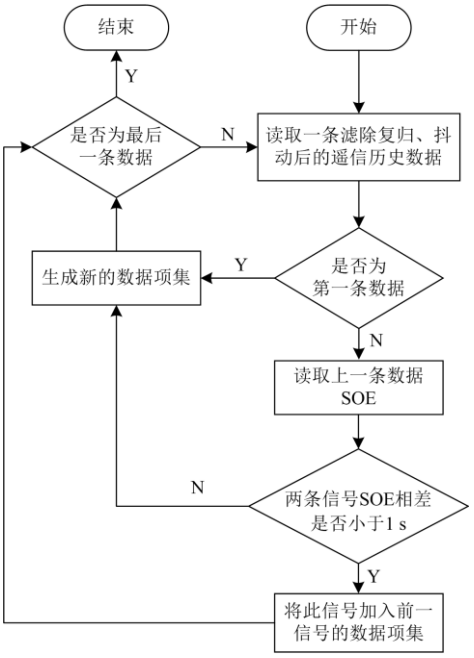

图 3 清洗后数据生成数据项集流程

Fig. 3 Process of generating from data cleaned to item dataset

表 1 某变电站某一时段遥信历史数据断面

Table 1 Data section of historical remote signals of a substation in some period of time

遥信名称	装置名称	遥信状态	遥信 SOE
服务器运行状态恢复	服务器 1	—	2019-01-31 04:47:09.254
服务器运行状态恢复	服务器 1	—	2019-01-31 04:47:09.271
出线控制回路异常	10 kV 出线 20 测控装置	动作	2019-01-31 20:44:25.140
母线保护启动	110 kV 母线保护装置	动作	2019-01-31 20:44:25.145
110 kV 测控装置闭锁	公共测控装置 1	动作	2019-01-31 20:44:25.278
110 kV 测控装置闭锁	公共测控装置 1	复归	2019-01-31 20:44:28.283
出线控制回路异常	10 kV 出线 20 测控装置	复归	2019-01-31 20:44:32.142
母线保护启动	110 kV 母线保护装置	复归	2019-01-31 20:44:33.262
母线保护录波就绪	110 kV 母线保护装置	动作	2019-01-31 20:44:33.363
母线保护录波就绪	110 kV 母线保护装置	复归	2019-01-31 20:44:37.808

其他遥信历史数据也可处理为类似的事务数据集，最终组合成一个事务数据集文本，作为数据挖掘的对象。

3 分布式并行 FP-growth 算法

3.1 Hadoop 框架

Hadoop 框架是 Apache 基金会开发的分布式系统基础架构，方便不熟悉分布式底层细节的用户开发分布式功能。Hadoop 主要由分布式存储系统、分布式计算系统(Mapreduce)、Hive、Hbase 等成员构成。其中，HDFS 和 Mapreduce 是最基础最重要的元素，HDFS 是系统的底层，用于存储集群中所有存储节点文件，Mapreduce 是一种编程模型，用于执行大规模数据集的并行运算。

Mapreduce 的计算由抽象编程映射接口 Map 和抽象编程简化接口 Reduce 两部分组成。Map 将挖掘任务分解，采用分治的思想，将任务发送给各节点，进行并行计算。数据来源于 HDFS 的数据块，数据通过键值对(key，value)标识，key 标识每行的偏移量，value 表示每条事务数据。Map 处理后，输出新的(key，value)键值，key 为项，value 表示这个项出现一次。Reduce 将挖掘的结果汇合，得到整体的挖掘结果。在计算过程中遇到的容错处理、任务调度、负载均衡均由 Mapreduce 负责，使用者不必关心这些细节。

3.2 FP-growth 算法

FP-growth 算法是继 Apriori 提出的一种挖掘频繁项集和数据之间强关联规则的方法，它构建了 FP 树并将数据集存储在 FP 树结构中来发现频繁项集或频繁项对。FP-growth 只需要遍历两次数据集就可以获得频繁项集，效率比 Apriori 算法更高。事务集 $D=\{D_1,D_2,\cdots,D_i\}$ 是一系列事务的集合，可以用来表述一系列集中发生的事件，在本文中即为转换后的事务数据集。

FP-growth 完成频繁模式发现的过程可以分为两个部分。

1) 构建事件 FP 树。第一次遍历事务数据集，导出频繁项大于 1 的项目集，获得支持度计数，并定义最小支持度计数(min_*sup*)和最小置信度(min_*conf*)，剔除小于 min_*sup* 和 min_*conf* 的项，然后将频繁项数据集按降序排列。创建 FP 树的根节点“null”，第二次扫描数据集，每个项集的项都按次序处理，并对每个项集创建一个分支。这样就形成事件的 FP 树。

2) 从事件 FP 树中挖掘频繁项集。对每个项集找到其条件模式基(Conditional Pattern Base, CPB)，递归调用树结构，删除小于最小支持度的项。如果最终形成单一路径的树结构，则列举所有组合；如果不是单一路径的树结构，则继续调用树结构，直至形成单一路径的树结构。

FP-growth 算法将发现频繁模式的问题转换成在较小的条件数据库中去搜索一些较短的模式，使用了最不频繁的项集作为后缀，显著地降低了遍历项集的资源开销。

3.3 改进并行 FP-growth 算法

单一变电站数据量小，对其进行频繁项挖掘不易获得有效的关联规则。针对变电站监控系统数据分散存储的特点，提出基于 Hadoop 的分布式并行 FP-growth 挖掘算法，拓展数据挖掘范围，有效利用各变电站端数据，提升海量数据的挖掘效率。在主站端设置一个主节点，各变电站监控服务器作为分布式挖掘的子节点。

分布式并行 FP-growth 算法计算过程如下。

1) 数据分区

各变电站监控数据原本就存储在各变电站，每个变电站的数据作为数据的一个分区。对所有分区的遥信历史数据表进行数据清洗，获得无复归、抖动等信号的数据表，再选取其中一个变电站将其清洗后的遥信历史数据表转换为以关键字为标识的事务数据集，并发送给其他计算节点，其中文关键字作为分布式并行数据预处理的参照。

2) 并行计算支持度(F_list)

通过 Hadoop Mapreduce 框架完成所有项的支持度计算，如图 4 所示。输入为变电站二次设备监测数据库转换后的事务数据集文本，统计每个事件在数据项集中出现的次数，设定最小支持度，去除小于最小支持度的事件，然后按降序重新排列频繁异常项集，得到 F_list 集合。

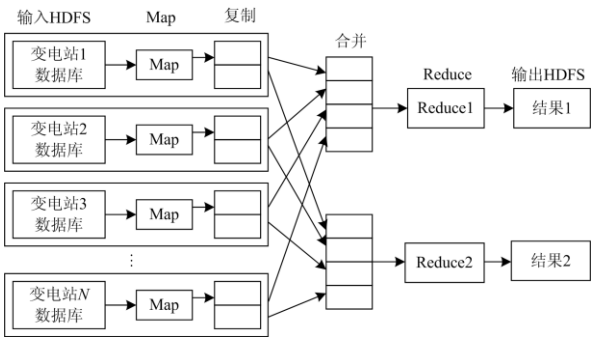

图 4 Mapreduce 计算框架

Fig. 4 Calculation framework of Mapreduce

全量扫描事务数据集文本，Map 函数将这些数据的键值对(key，value)标识转换为新的(key，value)键值，再存储到 HDFS。

3) item 分组

将 F_list 中的项分为 M 组，每组分配一个标识 group_id，所有事件及对应的 group_id 以 Hashmap 形式存储，形成表 Glist。

4) 并行 FP-growth 挖掘

并行 FP-growth 挖掘采用 Mapreduce 计算，Map 的输入来自各变电站数据库，为<key, value = 事件数据>。Map 同时读取 Glist 数据表。从事件数据最后一项依次向前扫描，若项 D_i 在 Glist 中对应的标识 group_id 第一次被扫描，则输出 $\{D_1, D_2, \cdots, D_i\}$，反之无输出。随后将同组别的事件数据输入一个 Reduce，组成一个 <key=group_id, value={事件数据 1}, {事件数据 2},…, {事件数据 N}>。每个 Reduce 就地构建 FP-tree，并将频繁模式放入支持度排序的大根堆中，输出支持度较高的频繁模式<key=事件项, value={包含该事件项的频繁模式}>。

5) 聚合生成频繁项集

主节点对各变电站子节点的挖掘结果进行聚集，选取不低于阈值的频繁项。当某子节点负载较重时，主节点可将挖掘任务向其他子节点平衡。主节点将挖掘结果发送给各子节点，各子节点根据频繁项名称进一步检索异常发生的时间、所属变电站名称、发生异常的设备名称、设备生产厂家、设备异常类型等其他维度的信息，形成文本或表格返回给主节点。主节点汇集各子节点返回的其他维度信息，再进一步综合分析产生异常信号的原因。

3.4 异常数据关联规则挖掘

关联规则的挖掘首先应检索项集来发现所有的频繁项集，再由频繁项集产生强关联规则。强关联规则是由最小支持度阈值 min_*sup* 和最小置信度阈值 min_*conf* 决定的。由于样本数量大，遥信信号类别多，各类异常样本的占有率低，因此选取最小支持度阈值 min_*sup*=0.05%，最小置信度阈值 min_*conf* 根据目标可分别设置为 60%，40%和 20%。

支持度和置信度的设置不足以滤掉负相关的关联规则，而强关联规则则可能湮没在繁杂的频繁项集中。因此文献[10]同时也引入了提升度的概念来判断关联规则是否具有有趣的关联关系，如式(2)，它表述的是事件 A 和 B 其中一个的出现提升了另一个出现的程度。

$$lift(\mathrm{A},\mathrm{B}) = P(\mathrm{A}\cup\mathrm{B}) / (P(\mathrm{A})P(\mathrm{B})) \tag{2}$$

式中：$P(A\cup B)$ 表示在 A 出现的情况下 B 事件发生的概率；$P(A)$ 表示事件 A 出现的概率；$P(B)$ 表示事件 B 出现的概率，提升度指标能够滤除具有误导性的负相关性的相关规则。

挖掘策略采用表 2 所列的模式，表 2 明确了挖掘的目标，并根据挖掘目标选取挖掘特征和维度选择特征。通过数据挖掘能够找到频发的异常名称、频发异常的设备，以及频发异常与设备之间的关系，支撑设备的故障诊断。首先采用频繁模式挖掘，挖掘出频繁异常信号，获得频发异常类型和日期，再进一步挖掘频发异常的设备名称。在此基础上挖掘频发异常和设备之间的关系，最后再根据频发日期挖掘频发信号之间的强关联关系。

表 2 异常数据挖掘策略

Table 2 Abnormal data mining strategy

序号	挖掘目标	挖掘特征	维度选择特征	挖掘方法
1	发现频发的异常名称	异常名称+异常类型	发现日期	频繁项挖掘
2	发现频发异常的设备	频发设备	异常名称+日期	频繁项挖掘
3	发现频发异常与设备之间的关联关系	异常名称+频发设备	发现日期	频繁项挖掘+关联规则挖掘

频繁项集内的元素之间不一定具有正相关关系，还需要进一步计算提升度，然后根据提升度来判断是否具有强关联关系。若 $lift(A,B)<1$，则表明 A 与 B 是负相关的；若 $lift(A,B)>1$，则表明 A 与 B 是正相关的，一个的出现则蕴含着另一个的出现；若 $lift(A,B)=1$，则表明 A 与 B 是独立的，两者之间没有关联性。

频繁项集的挖掘工作定期展开，每经过一段时间即对这一个时期内的历史数据进行清洗挖掘，形成新的 FP Tree，再生成新的频繁项集。这样就可以不断累积，在已挖掘数据的基础上对新产生的数据继续挖掘，节省计算资源。对异常信号的频繁项挖掘还能够发现设备的家族性缺陷，当同一类型设备、同一异常问题出现的频率超出一定次数时，有必要查找引起设备异常的原因，并组织专家评判设备是否存在家族性缺陷。

4　案例验证

为验证分布式并行 FP-growth 算法实际效果，搭建了一套实验环境，由 4 台 Hadoop 集群服务器构成，其中一台 64G 内存的服务器为主节点，其他 3 台 32G 内存的服务器视为变电站子节点。服务器均安装 Debian 9 x86 64 位操作系统，部署 Java 1.8.0 版和 MySQL 8.0 版。

挖掘的变电站自动化设备数据来源于 1 座 220 kV 和 2 座 110 kV 变电站二次设备监测系统 2018 年 8 月至 2019 年 9 月间的历史数据，约 80 万条数据，经过数据准备和清洗后，剩余 25 万条数据。分别采用分布式 FP-growth 算法和 Apriori 算法对三座变电站历史数据进行挖掘。FP-growth 算法挖掘的效率是 Apriori 算法的 3.4 倍。两种方法均挖掘出频繁异常事件 10 余项，包括加热器启动、公共测控温湿度控制器断线告警和 110 kV 保护测控装置 CPU 温度不确定状态等事件。其中一座 220 kV 变电站公共测控温湿度控制器断线告警出现的次数达 6 500 余次，加热器启动告警 2 420 次，两者频发的时间集中在 2019 年 1 月份。保护装置 CPU 温度不确定性告警出现 54 次。

经分析，两座变电站频繁出现某类型 110 kV 保护测控装置 CPU 温度不确定状态异常告警，两者为同型号装置，产生告警的原因是“110 kV 保护测控装置 CPU 温度不确定状态”信号的 Reference 是“LD001/STMP1STTmp”，此信号是量测类数据，Reference 应为“LD001/STMP1MXTmp”，而在此类型设备模型文件中却被误设为状态量，逻辑节点 STMP 的数据对象功能约束条件配置错误。经有现场运维经验的专家和厂家研发人员根据设备缺陷管理规定和监控信息处理管理规定确认，此类型装置的 ICD 模型文件配置错误，是家族性缺陷[25-26]，而测控装置温湿度控制器断线告警可能是由控制器负载不平衡引起的断线告警[27]。

当 min_*conf* 值设为 20%时，发现某变电站关联规则 1 条，即加热器启动与测控装置温湿度控制器断线两条告警存在关联性。两者的提升度 $lift(A,B)=3.28$，说明两者具有正的强相关性。这是一种单维关联规则，即先出现加热器启动告警，后出现测控装置温湿度控制器断线告警。

两者产生强关联规则关系的原因是温湿度控制器负载不平衡和整定值错误。两者频繁同时被触发的时间段集中在 2019 年 1 月前后，正是我国南方地区湿冷的季节，温湿度越限启动了温湿度控制器。温湿度控制器可接温度和湿度传感器两路输入信号，输出两路交流负载，而现场实施过程中，为适应本地设备运行要求，变电站内温湿度控制器是一个控制器带一路负载，且两路负载均使用相同的整定值，另一路空载。当环境温湿度达到整定值时，两路负载输出均被启动，空载的那路因温度控制器检测不到负载电路中的电流而报“温湿度控制器断

线告警”。

5 结束语

智能设备的运行状态关系着变电站能否安全稳定运行。本文提出了一种分布式并行挖掘智能变电站二次设备运行状态历史数据的方法，首先清洗复归、抖动等噪声信号，获得遥信事务数据集，采用 FP-growth 算法挖掘出频发设备异常和设备缺陷之间的关联关系，然后进一步分析关联告警之间的耦合关系。此方法在 Hadoop 集群服务器的实验证明，基于分布式并行 FP-growth 的挖掘方法能够有效发现设备自身的潜在缺陷、或设计配置方面存在的问题，为系统和设备的稳定运行清除了故障，也为二次设备家族性缺陷的认定提供数据基础。此方法不仅可以用于二次设备运维系统状态监测数据关联规则的挖掘，也可以用于一体化监控系统历史数据库数据的关联规则挖掘。

参考文献

[1] 叶远波，谢民，黄太贵，等. 基于灰色理论和云模型的智能变电站二次设备状态评估方法[J]. 电力系统保护与控制, 2019, 47(9):111-117.
YE Yuanbo, XIE Min, HUANG Taigui, et al. A state assessment method for intelligent substation secondary equipment based on grey theory and cloud model[J]. Power System Protection and Control, 2019, 47(9): 111-117.

[2] 代杰杰，滕莹冰，龚越明. 基于区间集聚类分析的电力设备状态异常检测方法[J]. 电力信息与通信技术, 2019, 17(11): 1-6.
DAI Jiejie, TENG Yingbing, GONG Yueming. Power equipment state anomaly detection method based on interval set theory and clustering analysis[J]. Electric Power ICT, 2019, 17(11): 1-6.

[3] 庄建煌，彭健，陈重，等. 基于泛在电力物联网的多技术融合变电设备发热缺陷判别预测系统的研究设计[J]. 高压电器, 2020, 56(9): 54-59.
ZHUANG Jianhuang, PENG Jian, CHEN Zhong, et al. Research and design of heating defect discrimination and prediction system for substation equipment based on multi-technology integration of ubiquitous power internet of things[J]. High Voltage Apparatus, 2020, 56(9): 54-59.

[4] 张立静，盛戈皞，江秀臣. 泛在电力物联网在变电站的应用分析与研究展望[J]. 高压电器，2020，56(9): 1-10.
ZHANG Lijing, SHENG Gehao, JIANG Xiuchen. Application analysis and research prospects of ubiquitous power internet of things in substation[J]. High Voltage Apparatus, 2020, 56(9): 1-10.

[5] 陈军，王利平，朱小红，等. 基于设备功能的智能变电站二次设备状态评估方法研究[J]. 电测与仪表，2019, 56(3): 47-51.
CHEN Jun, WANG Liping, ZHU Xiaohong, et al. State evaluation method of secondary device in smart substation based on function of device[J]. Electrical Measurement & Instrumentation, 2019, 56(3): 47-51.

[6] 张巧霞，王广民，李江林，等. 变电站远程运维平台设计与实现[J]. 电力系统保护与控制，2019，47(10): 164-172.
ZHANG Qiaoxia, WANG Guangmin, LI Jianglin, et al. Design and implementation of substation remote operation and maintenance platform[J]. Power System Protection and Control, 2019, 47(10): 164-172.

[7] 徐长宝，庄晨，蒋宏图. 智能变电站二次设备状态监测技术研究[J]. 电力系统保护与控制，2015，43(7): 127-131.
XU Changbao, ZHUANG Chen, JIANG Hongtu. Technical research of secondary equipment’s state monitoring in smart substation[J]. Power System Protection and Control, 2015, 43(7): 127-131.

[8] 戴志辉，张天宇，刘譞，等. 面向状态检修的智能变电站保护系统可靠性分析[J]. 电力系统保护与控制，2016, 44(16): 14-21.
DAI Zhihui, ZHANG Tianyu, LIU Xuan, et al. Research on smart substation protection system reliability for condition-based maintenance[J]. Power System Protection and Control, 2016, 44(16): 14-21.

[9] 笃骏，叶翔，葛立青，等. 智能变电站继电保护在线运维系统关键技术的研究与实现[J]. 电力自动化设备，2016, 36(7): 163-168.
DU Jun, YE Xiang, GE Liqing, et al. Key technologies of online maintenance system for relay protections in smart substation and its implementation[J]. Electric Power Automation Equipment, 2016, 36(7): 163-168.

[10] 韩家炜. 数据挖掘概念与技术[M]. 范明，孟小峰，等译. 北京：机械工业出版社, 2012.

[11] 宋亚奇，周国亮，朱永利. 智能电网大数据处理技术现状与挑战[J]. 电网技术, 2013, 37(4): 927-935.
SONG Yaqi, ZHOU Guoliang, ZHU Yongli. Present status and challenges of big data processing in smart grid[J]. Power System Technology, 2013, 37(4): 927-935.

[12] 王磊，陈青，高洪雨，等. 基于大数据挖掘技术的智能变电站故障追踪架构[J]. 电力系统自动化, 2018, 42(3): 84-91.
WANG Lei, CHEN Qing, GAO Hongyu, et al. Framework of fault trace for smart substation based on big data mining technology[J]. Automation of Electric Power Systems, 2018, 42(3): 84-91.

[13] KIRANMAI S A, LAXMI A J. Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy[J]. Protection and Control of Modern Power Systems, 2018, 3(3): 303-314. DOI: 10.1186/s41601-018-0103-3.

[14] 皇甫汉聪, 肖招娣. 基于 Apriori 算法的电力系统二次设备缺陷数据挖掘与分析研究[J]. 电子设计工程, 2019, 27(5): 12-15.
HUANGFU Hancong, XIAO Zhaodi. Data mining and analysis of two equipment defects based on Apriori algorithm[J]. Electronic Design Engineering, 2019, 27(5): 12-15.

[15] 周云祥. 基于数据挖掘的变电站监控后台告警信号自动分析[D]. 北京: 华北电力大学, 2016.
ZHOU Yunxiang. Automatic analysis of substation monitoring background alarm signals based on data mining[J]. Beijing: North China Electric Power University, 2016.

[16] 韩军, 马佳豪, 李晓军, 等. 基于数据挖掘的变电站告警数据分析和巡检策略研究[J]. 电力大数据, 2019, 22(7): 20-26.
HAN Jun, MA Jiahao, LI Xiaojun, et al. Research on substation alarm data analysis and patrol inspection strategy based on data mining[J]. Power System and Big Data, 2019, 22(7): 20-26.

[17] 张斌, 滕俊杰, 满毅. 改进的并行 FP-growth 算法在工业设备故障诊断中的应用研究[J]. 计算机科学, 2018, 45(6): 508-512.
ZHANG Bin, TENG Junjie, MAN Yi. Application research of improved parallel FP-growth algorithm in fault diagnosis of industrial equipment[J]. Computer Science, 2018, 45(6): 508-512.

[18] 李赞, 王朝霞, 孟月昊, 等. 基于 FP-growth 的前后部项约束关联规则改进算法[J]. 舰船电子工程, 2018, 38(9): 21-26.
LI Zan, WANG Zhaoxia, MENG Yuehao, et al. Improved algorithm of fore-part and rear-part item-constrained for mining association rules based on FP-growth[J]. Ship Electronic Engineering, 2018, 38(9): 21-26.

[19] 刘思怡, 苏运, 张焰. 基于 FP-Growth 算法的 10 kV 配电网分支断线故障诊断与定位方法[J]. 电网技术, 2019, 43(12): 4575-4581.
LIU Siyi, SU Yun, ZHANG Yan. Open-line fault diagnosis and positioning method for 10 kV power distribution network branch line based on FP-growth algorithm[J]. Power System Technology, 2019, 43(12): 4575-4581.

[20] 张延旭, 胡春潮, 黄曙, 等. 基于 Apriori 算法的二次设备缺陷数据挖掘与分析方法[J]. 电力系统自动化, 2017, 41(19): 1-5.
ZHANG Yanxu, HU Chunchao, HUANG Shu, et al. Apriori algorithm based data mining and analysis method for secondary device defects[J]. Automation of Electric Power Systems, 2017, 41(19): 1-5.

[21] 陈勇, 李胜男, 张丽, 等. 基于改进 Apriori 算法的智能变电站二次设备缺陷关联性分析[J]. 电力系统保护与控制, 2019, 47(20): 135-141.
CHEN Yong, LI Shengnan, ZHANG Li, et al. Association analysis for defect data of secondary device in smart substations based on improved Apriori algorithm[J]. Power System Protection and Control, 2019, 47(20): 135-141.

[22] 肖永立, 刘松, 见伟, 等. 一种基于 FP-growth 算法的变电站二次设备缺陷分析方法[J]. 电测与仪表, 2020, 57(12): 83-90.
XIAO Yongli, LIU Song, JIAN Wei, et al. A kind of defects analysis method for secondary device of substation based on FP-growth algorithm[J]. Electrical Measurement & Instrumentation, 2020, 57(12): 83-90.

[23] 马月坤, 刘鹏飞, 张振友, 等. 改进 FP-growth 算法及其分布式并行实现[J]. 哈尔滨理工大学学报, 2016, 21(2): 20-27.
MA Yuekun, LIU Pengfei, ZHANG Zhenyou, et al. Improved FP-growth algorithm and its distributed parallel implementation[J]. Journal of Harbin University of Science and Technology, 2016, 21(2): 20-27.

[24] 舒斐, 陈涛, 王斌, 等. 一种基于 DBN-BF 的电网工控系统异常识别方法[J]. 计算机工程, 2020, 46(11): 35-41.
SHU Fei, CHEN Tao, WANG Bin, et al. An abnormal identification method for smart grid industrial control system based on DBN-RF[J]. Computer Engineering, 2020, 46(11): 35-41.

[25] 国家电网公司运检部. 国家电网公司电网设备缺陷管理规定: 国网(运检/3)297—2014[S]. 北京: 国家电网公司运检部, 2014.

[26] 国家电网公司. 国家电网公司调控机构设备监控信息处置管理规定: (调/4)223—2014[S]. 北京: 国家电网公司, 2014.

[27] 周兴福, 韦良, 刘洋, 等. 变电站温湿度控制器断线报警分析[J]. 山东电力技术, 2016, 43(1): 74-76.
ZHOU Xingfu, WEI Liang, LIU Yang, et al. Disconnection alarm of temperature and humidity controller in transformer substations[J]. Shandong Electric Power, 2016, 43(1): 74-76.

收稿日期：2020-06-23；　修回日期：2020-11-11

作者简介：

方晓洁(1983—)，男，通信作者，硕士，高级工程师，研究方向为电力系统自动化；E-mail: xiaojie_fang@yeah.net

黄伟琼(1979—)，男，硕士，高级工程师，研究方向为电力系统自动化。E-mail: wqiong_huang@yeah.net

（编辑　姜新丽）