基于 FP-growth 算法的电压事件干扰源定位方法

许延祥[1]，曹军威[1]，许杏桃[2]，陈兵[2]，邓珂琳[1]

（1.清华大学 信息技术研究院，北京市海淀区 100084；2.江苏省电力公司 电力科学研究院，江苏省南京市帕威尔路 1 号 211103)

A Method of Locating Voltage Disturbance Sources Based on FP-growth Algorithm

Xu Yan-xiang[1], Cao Jun-wei[1],Xu Xing-tao[2], Chen Bing[2], Deng Ke-lin[1]

(1. Research Institute of Information Technology, Tsinghua University, Haidian District, Beijing, 100084, 2. State Grid Jiangsu Electric Power Compay Reserch Institute, Paweier Road No.1, Nanjing, Jiangsu Province, 211103)

ABSTRACT: Positioning disturbance sources is the precondition of confirming the liability for reducing power pollution in power system. This paper presents a method to analyze the mutual influence of the voltage events generated from the different nodes and then locate the voltage disturbance sources in a grid based on massive existing power quality monitoring data and association rule algorithm. This paper focuses on the implementation of transforming a batch of actual power quality data from a single-node sequential list into a multi-node parallel two-dimensional table. After that, we chose the PF-growth algorithm with some appropriate parameters to calculate the affecting relationship on voltage events between the nodes. Relative to the traditional methods, such as system simulation and matrix calculations, this method has low cost, fast and efficient computing features.

KEY WORDS: power quality, disturbance location, association rules algorithm, grid data mining

摘要：定位电能质量干扰源是确定电力系统中电能质量污染责任并深化治理的前提。本文提出一种利用已有的海量电能质量监测数据，基于关联分析算法确定各监测节点在电压事件上的相互影响关系，进而定位干扰源的方法。在技术路线上，本文重点实现了对采自实际电网的电能质量监测数据的转换处理，使顺序存储的数据转为多节点按时间轴对齐的二维表，之后选用 PF-growth 算法，采用合适的参数，计算出各节点之间在电压事件上的影响关系。相对于传统的系统仿真和矩阵计算的方法，本文方法具有成本低、计算快速有效等特点。

关键字：电能质量、干扰源定位、关联规则、电网数据挖掘

1.引言

电力系统中各种扰动引起的电能质量问题主要可分为稳态电能质量问题和暂态电能质量问题。稳态问题以波形畸变为特征，主要包括谐波、间谐波、噪声和频率波动等；暂态问题通常是以频谱和暂态持续时间为特征，可分为脉冲暂态和振荡暂态两类，主要包括电压跌落、电压骤升、短时断电和电容充电暂态等[1]。暂态问题中的电压骤降问题由于其发生的可能性远大于电压中断，即使几百公里以外的故障也有可能引起本地的电压跌落，因此在工业化国家，电压骤降已上升为最重要的电能质量问题之一，在国际上受到特别关注。国内电力企业对电压骤降的关注比其他问题电能质量问题的关注程度要高得多，同时用户对电压暂降引起问题的投诉占到全部投诉的 80%以上[2]。尤其是近 10 年来，随着高新技术、信息技术飞速发展，基于计算机控制的各种生产生活中的用电设备大量使用，对电压变化更敏感。随着电压骤降问题的日益突出，对电力系统中的电压骤降予以及时发现并判明干扰源头等工作显得尤为重要，可以为电力系统发现问题、明确问题责任并提高电能质量提供工作基础[3，4]。

在发现问题方面，主要依据时间轴上电压、电流、频率、相位等基本监测参数变化来确定是否发生了电压暂降问题[5,6]。监测工作主要依赖于各电子厂商生产的丰富电能质量监测仪器来完成，具有快速、准确的特点。但是，基于仪器只能判明某条线路上是否发生了电压暂降，而对于该电压暂降问题是本地线路所引起还是由相邻的其它线路所引起的问题，并不能给出有效的答案或判据[7]。

传统上对于电压事件干扰源的判定通常是基于电压事件监测数据，结合调度事件和电力系统的异常事件记录进行手工分析，具有定位精确、原因明确的优点，但同时也有耗时费力、主观性强及易受事件记录完备性影响等缺点[8]。

当前，电能质量监测系统已经开始广泛部署，部分省网公司已经把电能质量测量延伸到 10kv 线路，未来随着用电方对电能质量问题的愈加重视以及电能质量商品化的趋势驱动，电能质量监测系统必将延伸至用户进线部分。同时，计算科学中的数据挖掘方

法为分析电能质量问题带来新的技术手段。基于全网的同步电能质量监测数据，运用丰富的数据分析手段和数据挖掘方法，我们有望基于数据进行电压暂降干扰源的判断。

本文提出采用基于关联规则分析算法，面向电网局部区域化进行整体分析，从而快速定位电压暂降干扰源的判定，为高效率低成本地进行电能质量问题分析开拓一条新路径。

2.关联规则分析法

关联规则是挖掘数据库中两个或多个变量之间存在的隐含的关系。挖掘顾客交易数据库中项集间的关联规则问题由 Agrawal 等于 1993 年首先提出[9]，以后人们对关联规则挖掘问题进行了大量的研究。主要研究内容是对原有算法进行优化，如引入事务压缩、杂凑、数据库划分、随机采样、并行的思想等，以提高算法挖掘规则的效率；对关联规则的应用进行推广。

2.1 关联规则基本定义

定义 1 (关联规则) 关联规则是由 Agrawal 等人首先提出的一个重要 KDD 研究课题，它反映了大量数据中项目集之间有意义的关联或相关联系。

定义 2 (项) 设 $I = \{ i_1, i_2, \cdots, i_k\}$是一个二进制数字的集合，其中的元素称为项(item)。

定义 3 (支持度) 记 D 为交易(transaction) T 的集合，交易 T 是项的集合，并且 $T \subseteq I$。支持度 support $(A \Rightarrow B) = P(A \cup B)$ ，其中，$A \subset I$，$B \subset I$，并且 $A \cap B = \varnothing$。

定义 4 (置信度) $confidence(A \Rightarrow B) = P(B|A) = support(A \Rightarrow B)/support(A)$，其中，$A \subset I, B \subset I$ ，并且 $A \cap B = \varnothing$。

定义 5 (强关联规则) 是指挖掘出支持度大于客户指定的最小支持度(min_supp)和可信度大于最小可信度(min_conf)的关联规则。

定义 6 (频繁项集) 如果项集的出现频率大于或等于 min_supp 与 D 中事务总数的乘积，则称它为频繁项集。

2.2 关联规则评价及提升度

支持度是一个重要的度量，如果支持度很低，代表这个规则只是偶然出现，基本没有意义。因此，支持度通常用来删除那些无意义的规则。而置信度是通过规则进行的推理具有可靠性。对于规则 R:X⇒Y，只有置信度越高，Y 出现在包含 X 的事务中的概率才越大，否则这个规则也没有意义。

通常做关联规则会预设定最小支持度阈值 min_sup 和最小置信度阈值和 min_conf，而关联规则发现则是确定那些支持度大于等于 min_sup 并且置信度大于 min_conf 的所有规则（即“强关联规则”）。

单纯用支持度-置信度框架评价关联规则具有一定局限性。例如，如果图书市场中文学类书籍的数量远大于物理类书籍，那么物理类书籍的规则支持度就会很低，这样就导致很多物理类书籍的关联规则都被过滤掉了。再例如，如果 1000 个人中有 200 人喜欢喝茶，其中有 150 人喜欢喝咖啡，50 人不喜欢，那么我们通过置信度计算发现规则“R：喝茶⇒喝咖啡”的置信度非常高。但是可能另外不喜欢喝茶的 800 人中，有 650 人喜欢喝咖啡。由此可见喝茶和喝咖啡是两个独立事件，置信度度量忽略了规则后件中项集的支持度。

为了解决上述问题，引入了提升度(lift)的概念[]，来计算置信度和规则后件项集支持度的比率：

lift(A⇒B) = confidence(A⇒B) / support(B) =(p(A,B) / p(A)) / p(B)= p(A,B) / p(A)p(B)

lift(A⇒B)也称为兴趣因子，表示为 I(A,B)

通过概率学知识我们可以知道，如果 A 事件和 B 事件相互独立（或者我们称之为满足事件独立性假设），那么 p(A,B) = p(A) * p(B)，那么我们则可以这样来表示兴趣因子的度量：

当 I(A,B) = 1 时，我们称 A 和 B 是相互独立的，当 I(A,B) < 1 时，我们称 A 和 B 是负相关的，否则我们称 A 和 B 是正相关的。

2.3 关联规则的经典算法 Apriori 算法

挖掘顾客交易数据库中项集间的关联规则由 Agrawal 等于 1993 年首先提出，并设计了一个基本算法，其核心是基于频集理论的递推方法，即基于两阶段频集思想的方法，将关联规则的设计分解为两个子问题：1）. 找到满足最小支持度阈值的所有项集，我们称之为频繁项集。（例如频繁二项集，频繁三项集）；2）从频繁项集中找到满足最小置信度的所有规则。

由于步骤 2 中的操作极为简单，因此挖掘关联规则的整个性能就由步骤 1 中的操作处理所决定。挖掘关联规则的总体性能由第一步决定，第二步相对容易实现。首先产生频繁 1-项集 L_1，其次是频繁 2-项集 L_2，直到存在某个 r 值使得频繁项集 L_r为空，此时算法停止。其中在第 k 次循环中，产生候选 k-项集的集合 C_k，C_k中的每一个项集是由两个只有一个项不同的属于 L_{k-1}的频集通过与(k-2)-项集连接来产生的。

C_k中的项集是用来产生频集的候选集，其中最后频集 L_k必须是 C_k的一个子集。C_k中的每个元素需

在交易数据库中需要验证来决定其是否加入 L_k，这个验证过程是影响算法性能的一个瓶颈。可能产生大量的候选集，以及可能需要重复扫描数据库，是 Apriori 算法的两大缺点。

2.4 FP-growth 频集算法

针对 Apriori 算法的固有缺陷，J. Han 等提出了不产生候选挖掘频繁项集的方法：FP-growth 算法[10]。采用分而治之的策略，在经过第一遍扫描之后，把数据库中的频集压缩进一棵频繁模式树（FP-tree），同时依然保留其中的关联信息，随后再将 FP-tree 分化成一些条件库，每个库和一个长度为 1 的频集相关，然后再对这些条件库分别进行挖掘。当原始数据量很大的时候，也可以结合划分的方法,使得一个 FP-tree 可以放入主存中。实验表明，FP-growth 对不同长度的规则都有很好的适应性，同时在效率上较之 Apriori 算法有巨大的提高。具体算法分为两步：

（1）构造 FP-Tree

挖掘频繁模式前首先要构造 FP-Tree，算法如下：

输入：一个交易数据库 DB 和一个最小支持度 threshold。

输出：它的 FP-tree。

步骤：

1）扫描数据库一遍，得到频繁项的集合 F 和每个频繁项的支持度，把 F 按支持度递降排序，结果记为 L。

2）创建 FP-tree 的根节点，记为 T，并且标记为'null'，然后对数据库中的每个事务做如下的步骤：

根据 L 中的顺序，选出并排序 Trans 中的事务项，把 Trans 中排好序的事务项列表记为[p|P]，其中 p 是第一个元素，P 是列表的剩余部分。调用函数 insert_tree([p|P],T)，其运行如下：

如果 T 有一个子结点 N，其中 N.itemName=p.itemName，则将 N 的 count 域值增加 1；否则，创建一个新节点 N，使它的 count 为 1,使它的父节点为 T，并且使它的 nodeLink 和那些具有相同 itemName 域串起来。如果 P 非空,则递归调用 insert_tree(P,N)。

（2）挖掘频繁模式

对 FP-Tree 进行挖掘，算法如下：

输入：一棵用算法一建立的树 Tree

输出：所有的频繁集

步骤：

调用 FP-growth(Tree,null)。

```
procedure FP-Growth (Tree, x)
{
    if (Tree 只包含单路径 P) then
    {
        对路径 P 中节点的每个组合（记为 B）
        生成模式 B 并 x，支持数=B 中所有节点的最小支持度
    }else (对 Tree 头上的每个 ai，) do
    {
        生成模式 B= ai 并 x，支持度=ai.support;
        构造 B 的条件模式库和 B 的条件 FP 树 TreeB;
        if TreeB != 空集 then
            call FP-Growth (TreeB , B )
    }
}
```

3.电压暂降的关联规则分析建模

当前，电能质量监测仪器广泛部署，多省市已初步实现监测网络化。建有监测网的省网公司的数据中心经数年运行，已经积累了海量的电能质量问题监测数据。利用这些采集自各监测点的数据，我们有望通过比较同期电能质量问题记录，来判断不同监测节点相互之间的影响关系，从而确定电能质量干扰源所在线路或所在线路区域。

本文将基于关联分析算法定位干扰源的方法建模如下：

1）关联分析用于发现项集中项的关联，因而我们把每个监测点抽象为项集的一项，表现为二维数据表中的一列；

2）我们把所有电能质量监测节点在同一时刻上的同一种类的电能质量问题监测记录作为一个事务；

3）对于当前电能质量问题记录方式来讲，同一时刻同一问题的不同相位（A,B,C）的记录记为不同的事务；

4）对于谐波来讲，不同谐波记为不同的问题。

5）关联分析的结果以关联规则集合来体现，每条关联规则又体现项集之间的时间秩序关系（因果关系），在本模型中解释为节点（集）与节点（集）之间在某电能质量问题上的影响关系。

综上，本方法建模如图 1 所示。

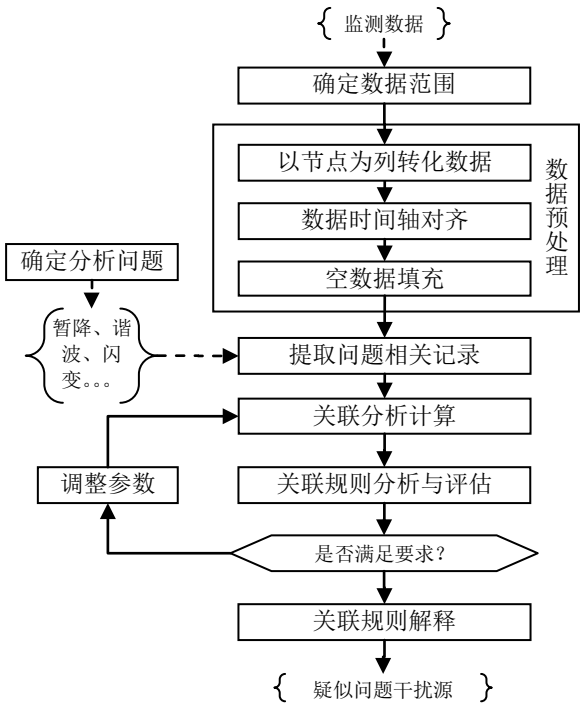

图 1 电能质量关联分析建模流程

Figure 1. Modeling Flow of PQ Association Analysis

4.原型系统的实现与分析

本文以江苏省电网的电能质量监测数据为基础，进行了基于关联分析的电压事件干扰源定位系统实现。

4.1 初始数据结构

本文实验数据来自江苏省电网监控中心的数据库，该数据采集全省 1000 多个监测点。本文用到的主要数据是电能质量历史数据超标表（dat_overrun），其基本结构如下图所示。

OID	CHV_OID	CHI_OID	DAT_OID	KIND	OC_DATE	SEQ	COL
1304382424	0	3131	6898101202	6	"2012-08-15 00:10:00"	"A"	"11"
1304382425	0	3131	6898101202	6	"2012-08-15 00:10:00"	"A"	"24"
1304382426	0	3131	6898101202	6	"2012-08-15 00:10:00"	"A"	"25"
1304382427	0	3131	6898101203	6	"2012-08-15 00:20:00"	"A"	"24"
1304382428	0	3131	6898101203	6	"2012-08-15 00:20:00"	"A"	"25"
1304382429	0	3131	6898101204	6	"2012-08-15 00:30:00"	"A"	"24"
1304382430	0	3131	6898101204	6	"2012-08-15 00:30:00"	"A"	"25"
1304382431	0	3131	6898101205	6	"2012-08-15 00:40:00"	"A"	"24"
1304382432	0	3131	6898101205	6	"2012-08-15 00:40:00"	"A"	"25"
1304382433	0	3131	6898101206	6	"2012-08-15 00:50:00"	"A"	"24"
1304382434	0	3131	6898101206	6	"2012-08-15 00:50:00"	"A"	"25"

图 2 电能质量历史数据超标表结构示意图

Figure 2 Structure of Historical PQ Overrun Data Table

其中，chi_oid 是线路监测点 ID，kind 代表电能质量问题类型，其中

1 代表：频率，

2 代表：短闪变，

3 代表：长闪变，

4 代表：电压不平衡，

5 代表：谐波电压，

6 代表：谐波电流，

7 代表：总谐波畸变率，

8 代表：电压变动，

9 代表：电压偏差'

oc_date 是记录时间，seq 代表出问题的相位，col 字段用于记录谐波波次。总计记录数据为 4232007 条。

另有电压通道表 DV_CHV、母线关联 DV_CHV_REF、电流通道表 DV_CHI、线路关联 DV_CHI_REF、变电站(监测网) DV_STATION、变电所关联 DV_STATION_REF 等系统级数据表用于记录各监测点在电网中的位置信息。

4.2 数据预处理

本文面向整个省网监测系统，因而首先从原始数据集中提取包含的所有监测节点（以下简称“节点”），以此生成新数据集的列。

原始数据集中，除谐波电流问题记录用 chi_oid 唯一标识一个节点外，其余电能质量问题均用 chv_oid 唯一标识一个节点。本文没有拿到进一步资料说明 chi_oid 与 chv_oid 是一一对应并同标识的，因而，本文将数据分成谐波电流和其它问题两个部分处理。新表数据结构如下：

表 1 预处理数据表结构

字段	类型	值域	说明
Time	datetime		时间点
Kind	Int	1-9	问题类型
Seq	Char	T,A,B,C	相位
Col	Int	2-50	谐波波次
Node_7	tinyint	0,1	节点 7 发生问题否
Node_8	tinyint	0,1	节点 8 发生问题否
…	tinyint	0,1	其它节点

生成两个数据表，其中谐波电流表（记为 chi_overrun）337 个字段，包含 333 个节点（项），其它问题表（记为 chv_overrun）544 个字段，包括 540 个节点。本文以 python 编程方式生成这两个表。

对于来自监测系统的原始数据，通过反复扫描数据表，将同一问题下同一时间点上所有结点的数据写入新数据表的同一条记录，对发生问题结点的字段记为 1，未发生相应问题的监测节点对应的列填入缺省数值 0。原始数据经数据转换后填入新表的情况如表 2 所示。

表 2 预处理后数据记录变化对比

Table 2. Record Count Comparison Between Preprocessed Data and Original Data

问题类别	原始记录数	转换后记录数
1 频率	846	179
2 短闪变	13380	3720
3 长闪变	12935	3793

4 电压不平衡	2666	1422
5 谐波电压	2653366	227463
6 谐波电流	1255229	197847
7 总谐波畸变率	57858	6500
8 电压变动	140975	1772
9 电压偏差	94752	6491

4.3 基于 FP-growth 算法进行挖掘

从转换预处理后的数据集 chv_overrun 中，提取关于电压变动的所有数据，去掉时间（Time)、问题（Kind)、相位（Seq）、波次（Col）列，存为 CSV 文件。

在 weka3.7 的 Explorer 组件中打开该文件，使用非监督学习中的属性过滤器 NumericToNominal 对数据值型数据（0，1）进行离散化处理。之后，我们选择 Associator 中的 FP-growth 算法进行节点关联性挖掘。

其中 FP-growth 涉及的参数包括如下[11]：

表 3 FP-growth 算法参数说明表

Table 2. Description of Parameters in FP-growth Algorithm

参数	含义	本文设置
Delta	每次迭代中最小支持度的增值幅度	5
findAllRulesForSupportLevel	是否要找出满足支持度的全部规则	否
lowerBoundMinSupport	支持度的最小值	5
maxNumberOfItems	项集的最大项数	2
metricType	是指定选择哪个量进行排序,weka 提供四种排序方法，0=confidence ，1=lift ， 2=leverage ，3=Conviction；	0
minMetric	指你选定的那个排序参数的那个最小值	0.5
numRulesToFind	给出要输出多少条规则	20
positiveIndex	正值的属性索引，对于密集数据的二元属性索引设为“正”，对于稀疏数据属性索引总是设为“2”	2
upperBoundMinSupport	支持度的最大值	80

由于相邻线路才会产生直接影响，监测系统中提供了各监测节点所在线路信息。根据各线路所属的变电站，我们可以重点计算变电站内各线路之间的影响，再计算相邻变电站之间的线路影响。这样可以极大地降低计算复杂度。基于这样的原则，我们选择了相邻近的 50 个监测节点的数据进行分析。本文所示例的电能质量问题为电压偏差。

4.4 规则解释

经挖掘后，共产生 387 条关联规则（如下图所示）。

去除（0==>0, 0==>1, 1==>0）这样的非兴趣规则 115 条，通过对剩余 272 条规则进行分析，我们得出以下 2 条规律：

1）节点编号为 102、103、104、105、106、107、111、112、113、114、160、162、164 的 13 个节点在电压偏差问题具有较明显的共现关系，因而在其中一点发生电压偏差问题时，应向其它节点发出警报，或自动启动其它节点的防范动作；

2）节点编号为 103、104、105、106、107、111、112、113、114、158、159、160、161、162、163、164 等 16 个节点的电压偏差问题对节点 102 有明显的影响作用。

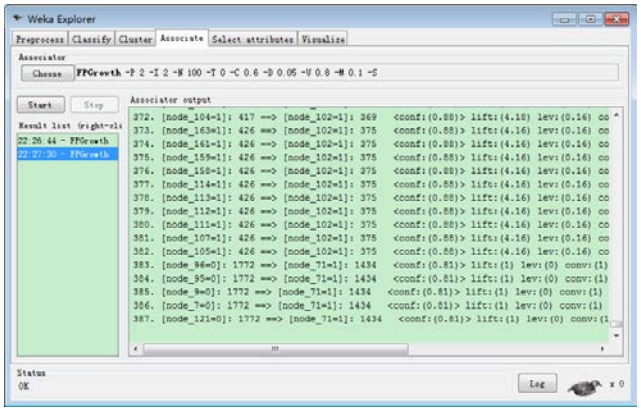

图 3 FP-growth 挖掘结果示意图

Figure 3. Illustration of Mining Result of FP-growth

5 总结

本文通过对阶段性的历史监测数据的统计分析，能够快速有效地定位出谐波源的所在范围，同时该方法并未对实际运行电网产生任何额外干扰。

随着电能质量监测网络的深入发展，电能质量监测仪必然更加广泛深入地部署到具体负荷和供电设备上。届时，将本文方法在更深入范围的监测数据上计算应用，将可精确确定产生电能干扰源的具体线路或设备。

1、 林海雪.现代电能质量的基本问题.电网技术, 2001, 25(10):5-12

2、 韩英铎,严干贵,姜齐荣等.信息电力与 FACTS 及 DFACTS 技术.电力系统自动化,2000(19):1-7

3、 肖湘宁,徐永海.电能质量问题剖析.电网技术,2001,25(3):66-69

4、 徐永海 ,肖湘宁.电力市场环境下的电能质量问题. 电网技术, 2004,28(22):48-52

5、 赵凤展,杨仁刚.基于短时傅里叶变换的电压暂降扰动检测[J].中国电机工程学报,2007,27(10):28 一 34,109.

6、 杨洪耕,刘守亮,肖先勇,等.基于 s 变换的电压凹陷分类专家系统[J].中国电机工程学报,2007,27(l):98 一 104.

7、 欧阳森.低压配电系统中电能质量监测的信号处理方法.[西安交通大学博士学位论文].西安:西安交通大学电气工程系,2003,1-2

8、 杨洪耕,刘守亮,肖先勇,等.基于 s 变换的电压凹陷分类专家系统[J].中

国电机工程学报,2007,27(l):98 一 104.
9、 Han J W,Kamber M.Data mining: concepts and techniques[M].Morgan Kaufmann,2005.
10、 Han jia wei, Pei Jan 等 Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach.2004
11、 J. Han, J.Pei, Y. Yin: Mining frequent patterns without candidate generation. In: Proceedings of the 2000 ACM-SIGMID International Conference on Management of Data, 1-12, 2000.

收稿日期：

作者简介：

许延祥(1975-)：男，博士，博士后，研究方向为面向电力系统的数据挖掘技术，fighter1975@163.com;

曹军威(1973-)：男，博士，研究员，研究方向为智能电网、能源互联网等;

许杏桃(1967-)：男，高级工程师，研究方向为无功电压优化控制和电能质量治理。