文章编号：1005-8451（2021）03-0059-06

基于改进 Apriori 算法的铁路网络安全预警方法研究

崔伟健[1]，马小宁[2]，孙思齐[2]
（1. 中国铁道科学研究院，北京 100081；
2. 中国铁道科学研究院 电子计算技术研究所，北京 100081）

摘 要：随着大数据技术的飞速发展，基于大数据技术的安全隐患事前感知技术日趋成熟，为分析海量数据和挖掘事故规律提供了有效手段。文章优化、改进数据挖掘技术中经典的 Apriori 算法，基于改进的 Apriori 算法，研究提出一种铁路网络安全预警方法，并结合网络安全等级保护要求，构建铁路网络安全指标体系，通过仿真实验，验证算法的正确性和可用性。研究表明，采用改进的 Apriori 算法能够实现铁路网络安全事前感知，有效解决预警防范手段不足的问题，在铁路网络安全预警中具有较高的应用价值。

关键词：网络安全；安全预警；数据挖掘；态势感知；等级保护
中图分类号：U29 : TP393 **文献标识码：**A

Railway network security early warning method based on improved Apriori algorithm

CUI Weijian[1], MA Xiaoning[2], SUN Siqi[2]
(1. China Academy of Railway Sciences, Beijing 100081, China；2. Institute of Computing Technologies, China Academy of Railway Sciences Corporation Limited, Beijing 100081, China)

Abstract: With the rapid development of big data technology, security hidden risk prior awareness technology based on big data technology is becoming more and more mature, which provides an effective means for analyzing massive data and mining accident rules. This paper optimized and improved the classical Apriori algorithm in data mining technology, based on the improved Apriori algorithm, proposed a railway network security early warning method, and combined with the requirements of network security level protection, constructed the railway network security index system. The correctness and availability of the algorithm were verified by simulation experiments. The research shows that the improved Apriori algorithm can implement the railway network security hidden risk prior awareness and effectively solve the problem of the lack of early warning and prevention means, which has a certain application value in railway network security early warning.

Keywords: network security；security early warning；data mining；situation awareness；classified protection

随着铁路信息化的快速发展，铁路各业务系统已高度依赖信息技术，随之而来的网络安全形势也日趋严峻，从简单的网络安全攻击试探、网页挂马篡改，到有组织、大规模的 DDos 攻击、APT（Advanced Persistent Threat）攻击等，对铁路信息系统稳定运行和各项业务稳定开展带来了严重威胁，甚至会影响铁路行车安全和人民生命财产安全[1]。近年来，铁路大力发展网络安全建设，初步建成网络安全防护体系，但网络安全建设工作起步较晚，目前，安全防护仍停留在常规保障、事后处理阶段，尚未建成行之有效的行业网络安全预警体系。

在食品卫生、公路交通、电网电力等行业，已开展了网络安全预警技术研究[2-4]，还有一些学者对自修正系数、异常流量分析、DDos 攻击分析等网络安全预警方法开展研究[5-7]，在铁路网络安全领域，针对系统、流量的网络安全预警方法研究相对较多，如基于网闸技术的高速铁路地震预警[8]、车货实时追踪预警[9]、基于大数据技术的网络安全态势感知[10]等，但从铁路网络安全合规测评角度开展网络安全预警的研究相对较少，在铁路大力发展网络安全测评和大数据建设的背景下，结合网络安全测评工作实际，

收稿日期：2020-10-26
基金项目：中国国家铁路集团有限公司重大课题（K2019S002）
作者简介：崔伟健，在读硕士研究生；马小宁，研究员。

研究基于大数据技术的网络安全预警方法具有较强的现实意义。

目前，Apriori 算法在铁路网络安全预测与感知[11-12]中应用较为广泛，且应用效果已得到广泛认可。本文分析了 Apriori 算法特性，提出了一种改进的 Apriori 算法，结合铁路某单位实际案例，对算法应用进行验证。研究表明，改进的 Apriori 算法在铁路网络安全预警工作中有较高的应用价值。

1 Apriori 算法及改进

1.1 术语定义

下面给出 Apriori 算法中涉及的部分术语定义。

（1）支持度和最小支持度：初始数据中包含某一项集的比例，用 $S\%$ 表示；在计算中，需满足的支持度的最小值记为最小支持度，用 min_Sup 表示。

（2）置信度和最小置信度：在迭代过程中，当前项集与其前一层相关子项集支持度的比例，用 $C\%$ 表示；在计算中，需满足的置信度的最小值记为最小置信度，用 min_Con 表示。

（3）候选项集：在迭代过程中，包含当前层所有项集的集合，用 $\boldsymbol{C_k}$ 表示。

（4）频繁项集：在迭代过程中，满足最小支持度项集的集合，用 $\boldsymbol{L_k}$ 表示。

（5）最大频繁项集：如果频繁项集 $\boldsymbol{L_k}$ 的所有超集都是非频繁项集，那么当前频繁项集 $\boldsymbol{L_k}$ 即为最大频繁项集，用 $\boldsymbol{MFI}$ 表示。

1.2 Apriori 算法

Apriori 算法核心是逐层搜索迭代，在每一层迭代中，由候选项集 $\boldsymbol{C_k}$（k=1,2,⋯,n）生成频繁项集 $\boldsymbol{L_k}$，在层间迭代中，由当前层频繁项集 $\boldsymbol{L_k}$ 生成下一层候选项集 $\boldsymbol{C_{k+1}}$，通过有限次数的迭代后，找出数据中隐含的最大频繁项集。

设 $\boldsymbol{I}=\{i_1,i_2,\cdots,i_m\}$ 为所有项目的集合，$\boldsymbol{T}$ 是一个由项目组成的集合，且满足 $\boldsymbol{T}\subseteq\boldsymbol{I}$；项目数据库 $\boldsymbol{D}$ 为 $\boldsymbol{T}$ 的集合，满足 $\boldsymbol{D}=\{\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_n\}$，$|\boldsymbol{D}|$ 是 $\boldsymbol{D}$ 的总项目数。设 $\boldsymbol{X}$，$\boldsymbol{Y}$ 为 $\boldsymbol{I}$ 中项的集合，满足 $\boldsymbol{X}\subset\boldsymbol{I},\boldsymbol{Y}\subset\boldsymbol{I}$，关联规则就是形如 $\boldsymbol{X}\Rightarrow\boldsymbol{Y}$ 的逻辑蕴含关系，且 $\boldsymbol{X}\cap\boldsymbol{Y}=\boldsymbol{\phi}$。

根据上述可以得到一个关联规则 $\boldsymbol{X}\Rightarrow\boldsymbol{Y}(S\%,C\%)$，其中，$S\%$ 为满足条件的项目占总项目数 $|\boldsymbol{D}|$ 的比例，即支持度，计算公式如式（1）：

$$Support(\boldsymbol{X}\Rightarrow\boldsymbol{Y})=\mathrm{S}\%=\frac{|\{\boldsymbol{T}:\boldsymbol{X}\cup\boldsymbol{Y}\subseteq\boldsymbol{T},\boldsymbol{T}\in\boldsymbol{D}\}|}{|\boldsymbol{D}|}\quad(1)$$

$C\%$ 为 $\boldsymbol{D}$ 中包含 $\boldsymbol{X}$ 项目又包含 $\boldsymbol{Y}$ 项目的比例，即置信度，计算公式如式（2）：

$$Confidence(\boldsymbol{X}\Rightarrow\boldsymbol{Y})=C\%=\frac{|\{\boldsymbol{T}:\boldsymbol{X}\cup\boldsymbol{Y}\subseteq\boldsymbol{T},\boldsymbol{T}\in\boldsymbol{D}\}|}{|\{\boldsymbol{T}:\boldsymbol{X}\subseteq\boldsymbol{T},\boldsymbol{T}\in\boldsymbol{D}\}|}\quad(2)$$

Apriori 算法流程如图 1 所示。

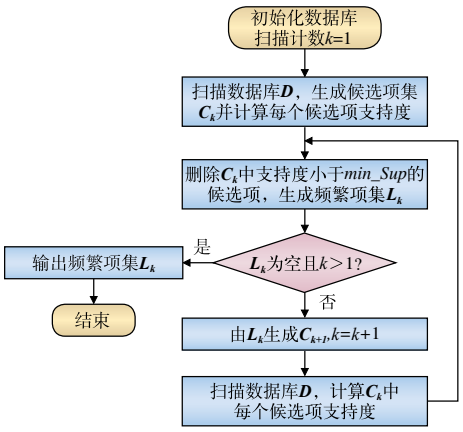

图1 Apriori 算法流程

Apriori 算法虽然应用广泛，但每层迭代计算都需对数据库进行多次扫描，数据量级大时存在数据库扫描频率高、算法效率低、会产生大量中间项集等不足。

1.3 Apriori 算法改进思路

许多学者对 Apriori 算法优化进行了研究，如张雷[13]等人在基于改进 Apriori 算法的客户需求数据分析方法中，提出基于布尔矩阵的改进 Apriori 算法，胡世昌[14]等人研究提出基于二进制编码的改进 Apriori 算法，殷茗[15]等人提出基于邻接表的改进 Apriori 算法，陈江平[16]等人提出利用概率方法改进的 Apriori 算法。各类改进的 Apriori 算法均实现了降耗提效，取得了一定效果。

本文从最大频繁项集的性质着手，研究提出一种 Apriori 算法改进方法。根据传统 Apriori 算法，最大频繁项集必定存在下述 3 个性质：

（1）最大频繁项集必为 N 个事务集的交集，且 $N\geqslant min\_Sup\_N$，其中，min_Sup_N 为最小支持数，其

值为事务集个数与最小支持度的乘积向上取整；

（2）最大频繁项集支持度一定小于或等于其他频繁项集支持度；

（3）最大频繁项集中所含元素项个数必定大于或等于其他频繁项集中元素项个数。

根据上述性质，可以得出以下 2 个推论：

（1）最大频繁项集可由 *min_Sup_N* 个事务集求交集得出；

（2）任意 *min_Sup_N* 个事务集求交集结果中，所含元素项个数最多的项集一定是最大频繁项集。

根据上述推论，可以得出 Apriori 算法改进思路，即通过求交集运算方式找出数据中的最大频繁项集。

1.4　改进的 Apriori 算法

预设项目集合 $\boldsymbol{I}$，事务集合 $\boldsymbol{T}$，项目数据库 $\boldsymbol{D}$，最小支持度 *min_Sup* 并计算出 *min_Sup_N*，则单次交集计算公式如式（3）：

$$\boldsymbol{Q_m} = intersection\left(\boldsymbol{T_i}, \boldsymbol{T_j}, \boldsymbol{T_k}, \cdots\right) \tag{3}$$

式（3）中，*intersection* 表示求交集运算，$\boldsymbol{T_i}$，$\boldsymbol{T_j}$，$\boldsymbol{T_k}$，…为参与计算的事务集，满足 $i \neq j \neq k \neq \cdots$，总数为 *min_Sup_N* 个，$\boldsymbol{Q_m}$ 为所求交集结果，通过不断交集计算，找出含元素项个数最多的 $\boldsymbol{Q_m}$，或 $\boldsymbol{Q_m}$ 的集合，即为所求最大频繁项集。

改进的 Apriori 算法流程如图 2 所示。

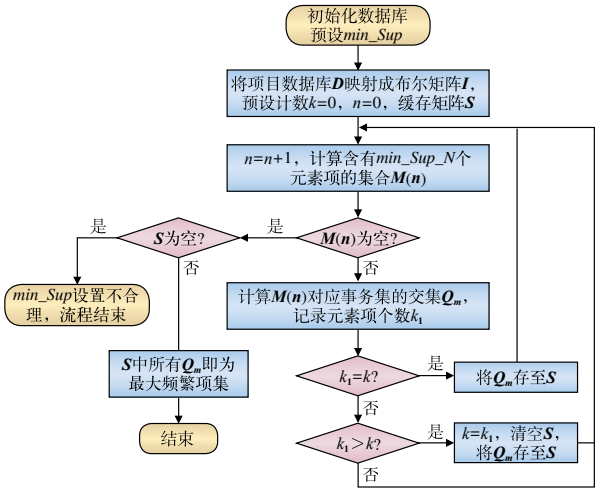

图2　改进的 Apriori 算法流程

2　改进的 Apriori 算法性能分析

与常规的正向推演、过程优化的改进方式不同，本文从最大频繁项集性质入手，对算法运算过程进行改进，与传统 Apriori 算法相比，优化了逐层迭代过程，在数据库扫描频率、计算复杂度、临时存储空间占用等方面有明显改善。

2.1　数据库扫描频率和计算稳定度分析

在数据库扫描方面，传统 Apriori 算法每层迭代都要遍历数据库，而改进的 Apriori 算法将原始数据映射至缓存矩阵中，在计算过程中，仅遍历一次数据库即可，数据库扫描频率显著降低，数据库 I/O 耗能明显减少。

在计算稳定度方面，传统 Apriori 算法在计算过程中会不断剪枝，因此，计算稳定度与初始矩阵维数、初始数据稠密程度、剪枝条件关系较大，而改进的 Apriori 算法计算量仅与初始数据维数、最小支持度有关，算法计算稳定度相对较高。

2.2　计算时间和复杂度分析

设 n 为事务集 $\boldsymbol{T}$ 的数量，m 为每个事务集 $\boldsymbol{T}$ 的平均项目数，q 为满足 *min_Sup* 的频繁项集元素个数最大值，t_D 为扫描初始数据元素项所需时间，t_a 为扫描数组每个元素所需时间，t_c 为每次进行简单四则运算所需时间，$t_{apriori}$ 和 $t_{new\_apriori}$ 分别代表传统和改进后 Apriori 算法计算出满足 *min_Sup* 的频繁项集所需时间。

（1）$t_{apriori}$ 时间性能分析

对于 $t_{apriori}$，由于传统 Apriori 算法在运算过程中不断进行剪枝操作，对初始数据较为敏感，初始数据矩阵较为稀疏时，逐层迭代过程中剪枝量相对较大，到下一个迭代层，运算量较低。而初始数据矩阵较为稠密时，其剪枝量相对较小，此时算法运算量在前几层迭代中会逐层递加，最终总运算量也会相对较大，传统 Apriori 算法的计算复杂度不稳定性在后续仿真实验中也进行了验证。另外，传统 Apriori 算法对 *min_Sup* 也比较敏感，*min_Sup* 越大，算法迭代次数越多，数据库 I/O 次数越多，消耗的 t_D 及运算次数也相对越大。

（2）$t_{new\_apriori}$ 时间性能分析

对于 $t_{new\_apriori}$，需要计算 $\frac{n!}{min\_Sup\_N! \cdot (n-min\_Sup\_N)!}$ 次交集，由于项目矩阵行元素项均为顺序排列，且对应元素项在同一列，故在计算 *min_Sup_N* 个矩阵

交集时，仅需对各矩阵对应列元素项进行求与操作即可，结果为 1，证明该元素项为交集元素；为 0，则证明不是交集元素。最终计算次数近似为 $(min\_Sup\_N-1)\cdot n$，因此，$t_{new\_apriori}$ 的计算公式如式（4）：

$$t_{new\_apriori}=\frac{m!}{min\_Sup\_N!\cdot(m-min\_Sup\_N)!}\cdot n\cdot\frac{}{(min\_Sup\_N-1)\cdot(t_c+t_a)} \quad (4)$$

从式（4）中可以看出，$t_{new\_apriori}$ 与 m、min_Sup_N 关联度较大，在实际应用过程中，m 的值在某一段时期内是固定的，因此，$t_{new\_apriori}$ 与 min_Sup_N 关联度最大，且为正相关关系。

综合来看，改进的 Apriori 算法更加稳定，在 n、m、min_Sup 值较大时，改进算法综合性能更优。

2.3 空间性能分析

传统 Apriori 算法对存储空间需求非常大，若初始元素个数为 10000，在第 2 轮迭代中，生成的含 2 个元素项的候选项集个数会接近 10^8 个，虽然中间会进行部分裁剪，但依然会消耗大量存储空间。改进的 Apriori 算法对空间需求有限，仅需预设一个映射矩阵 $\boldsymbol{I}$，以及一个缓存矩阵 $\boldsymbol{S}$，所需存储空间与中间频繁项集数量有关，但由于产生的中间频繁项集数量相对较少，故所需空间也较小。因此，改进的 Apriori 算法在空间性能上明显占优。

2.4 仿真验证

为检验上述算法分析的合理性，本文利用 Matlab 从两个方面进行了仿真验证。

（1）对改进前后算法的运算稳定性进行验证，预设 50 个项目集，每个项目集包含事务集数为 100，项目数为 100 个，min_Sup=20%，采用改进前后的 Apriori 算法计算最大频繁项集，传统 Apriori 算法对初始数据敏感度较高，计算复杂度因原始数据不同而不断变化，运算不够稳定，而改进的 Apriori 算法对初始数据不敏感，运算稳定度相对更高。计算结果如图 3 所示。

（2）进一步对改进前后算法的计算量进行分析，同样预设 50 个项目集，每个项目集包含事务集数的值分别为 60、70、80、90、100、110、120、130、140、150，项目数为 100 个，min_Sup=20%，考虑到

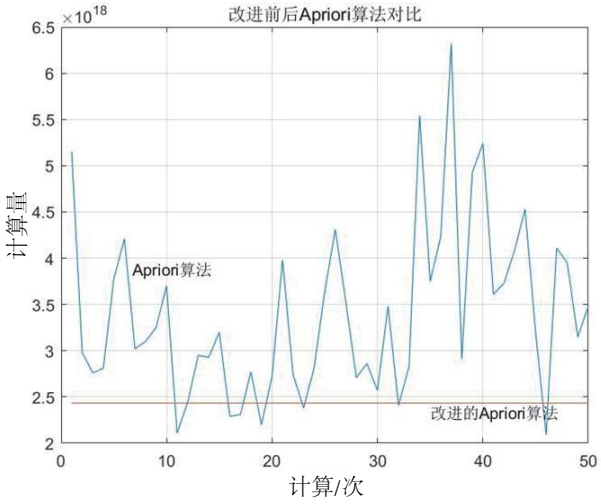

图3 改进前后 Apriori 算法计算稳定度对比

传统 Apriori 算法对初始数据较为敏感，用某个固定项目集进行验证可能出现较大误差，因此，对每个项目集取 100 组初始化数据，分别计算两种算法平均运算次数，虽然这种计算方式也存在一定误差，但由曲线趋势可以判断出，改进的 Apriori 算法计算性能明显更优。计算结果如图 4 所示。

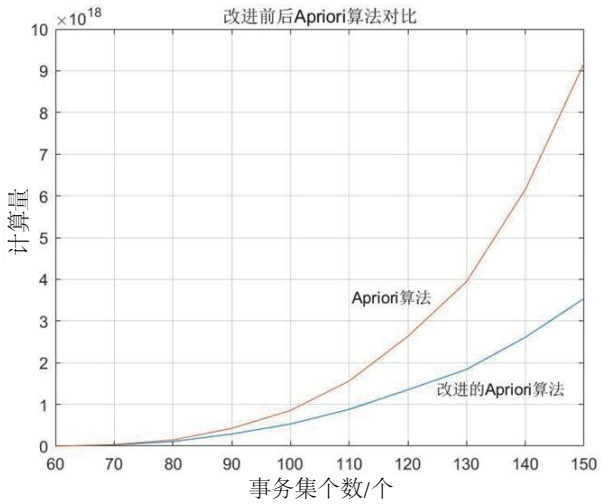

图4 改进前后 Apriori 算法计算次数对比

3 案例分析

本文结合国家等级保护 2.0 标准，构建了铁路网络安全指标体系[17]，共包括 211 项指标。各项指标的量化均根据实际评估结果得出。测评人员在对铁路信息系统安全检测评估过程中，对照具体指标项，核查指标符合情况，分为符合和不符合两项，打分标准对应为 1 分和 0 分。

以铁路某单位网络安全预警为例，选取 2019—2020 年该单位发生网络安全事件时系统各指标数据，从数据库中选取 300 组 211 项三级指标数据，基于改进的 Apriori 算法，利用 Matlab 进行仿真验证，挖掘得到网络安全风险强关联因素，由此给出网络安全预警及未来一段时间内网络安全重点防护方向。部分抽样出来的风险指标数据如表 1 所示。

表1　铁路某单位网络安全指标

标识符 TID	指标项										
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	…
1	0	2	3	4	5	6	7	8	0	0	
2	0	2	0	0	0	6	0	0	9	10	
3	1	2	0	0	0	6	0	8	9	10	
4	0	2	0	4	0	6	7	0	9	0	
5	0	0	0	0	5	0	7	8	0	10	…
6	1	2	3	4	0	6	7	8	9	0	
7	0	2	3	0	0	0	0	8	0	0	
8	0	2	0	0	5	6	0	8	9	10	
…											

由于篇幅有限，上表仅列出 M1 ~ M10 指标项量化后矩阵，其余 201 项指标量化数据未详细展示。这里简要介绍 M1 ~ M6 这 6 项指标含义。

M1：机房场地应具备防震、防风和防雨等能力；

M2：机房应具备防潮、防水措施；

M3：应设置电子门禁，记录、鉴别和控制进出人员；

M4：应固定设备主要部件，设置明显标识；

M5：通信线缆应在隐蔽安全处；

M6：机房应设置防盗措施。

设 *min_Sup*=20%、*min_Con*=50% 时，得出的结果如下：

$$\boldsymbol{L}=\begin{bmatrix} 18 & 24 & 39 & 49 & 60 & 192 \\ 24 & 56 & 60 & 94 & 109 & 115 \\ 24 & 56 & 94 & 98 & 109 & 192 \\ 24 & 56 & 96 & 109 & 166 & 182 \\ 24 & 56 & 96 & 109 & 192 & 201 \end{bmatrix}$$

各频繁项集对应的置信度分别为 0.63、0.50、0.50、0.56、0.50。

得出的最大频繁项集有 5 个，以第 1 个为例，得出网络安全事件与第 18、24、39、49、60、192 指标关联度较大，满足预设的最小支持度，且这 6 项指标同时不满足要求的情况下，发生网络安全事件的概率为 63%。说明该单位应着重强化安全管理制度制订、恶意代码防护、系统漏洞修复和补丁升级、系统应急预案制订和应急演练开展、网络安全专业技术人员配备等工作，加强网络安全建设和问题整改，最大程度避免网络安全事件的发生。

4　结束语

本文研究提出了一种铁路网络安全预警方法，结合铁路大数据实际应用情况，引入 Apriori 算法支持关联规则计算，针对传统 Apriori 算法在计算复杂、空间耗费高等方面的不足，对算法过程进行了优化改进，并通过理论分析和示例仿真验证了改进效果。最后结合铁路某单位实际案例，对算法在铁路网络安全预警方面的应用进行了分析验证。理论论证和实验结果表明，改进的 Apriori 算法性能良好，计算复杂度较低，在铁路网络安全预警领域具有一定的应用价值。

改进的 Apriori 算法在交集计算、过程剪枝等步骤中均存在一定的提升空间，实际计算过程中，大多数求交集计算属无意义计算，并不能得出有价值的结果，若对初始数据进行一定的质量分析，或者运算过程中合理设置相应策略，可进一步减少求交集次数，加大剪枝量，大幅度提升算法效率，这也将是后期工作的研究重点。

参考文献

[1] 张伯驹. 新形势下铁路网络安全工作探索与发展展望 [J]. 铁路计算机应用，2020，29（8）：1-5.

[2] 顾小林，张大为，张　可，等. 基于关联规则挖掘的食品安全信息预警模型 [J]. 软科学，2011，25（11）：136-141.

[3] 郭骁炜，樊升印. 网络安全态势感知在高速公路联网收费系统中的应用 [J]. 筑路机械与施工机械化，2020，37（3）：72-76.

[4] 刘红军，管　美，刘　勇，等. 电网调度系统网络安全态势感知研究 [J]. 电测与仪表，2019，56（17）：69-75.

[5] 杨宏宇，张旭高. 基于自修正系数修匀法的网络安全态势预测 [J]. 通信学报，2020，41（5）：196-204.

[6] 黎佳玥，赵　波，李　想，等. 基于深度学习的网络流量异常预测方法 [J]. 计算机工程与应用，2020，56（6）：39-50.

[7] 罗逸涵，程杰仁，唐湘滟，等. 基于自适应阈值的DDoS攻击态势预警模型 [J]. 浙江大学学报（工学版），2020，54（4）：704-711.

[8] 李京生，张 浞，赵 林. 基于网闸技术的高速铁路地震预警监测系统安全性研究 [J]. 铁道运输与经济，2018，40（7）：101-104.

[9] 关梦园，钱 琳，王钰鹏，等. 铁路车货实时追踪及预警系统关键技术研究 [J]. 科技成果管理与研究，2020，15（9）：39-42.

[10] 董 鹏，马小宁，高明星. 铁路网络安全态势感知平台方案研究 [J]. 铁路计算机应用，2020，29（4）：50-54.

[11] 万 斌，徐 明. 一种基于Apriori算法的网络安全预测方法 [J]. 电力信息与通信技术，2019，17（1）：133-138.

[12] 陆江东，郑 奋，戴卓臣. 基于改进Apriori的网络安全感知方法 [J]. 计算机测量与控制，2017，25（10）：244-246, 254.

[13] 张 雷，董万富，阚欢迎，等. 基于改进Apriori算法的客户需求数据分析方法 [J]. 机械设计与制造，2020（5）：185-188.

[14] 胡世昌，李劲华，王常颖. 基于二进制编码的Apriori改进算法 [J]. 计算机应用研究，2020，37（2）：398-400, 423.

[15] 殷 茗，王文杰，张煊宇，等. 一种基于邻接表的最大频繁项集挖掘算法 [J]. 电子与信息学报，2019，41（8）：2009-2016.

[16] 陈江平，傅仲良，徐志红. 一种Apriori的改进算法 [J]. 武汉大学学报（信息科学版），2003，28（1）：94-99.

[17] 刘 刚，杨轶杰. 基于等级保护2.0的铁路网络安全技术防护体系研究 [J]. 铁路计算机应用，2020，29（8）：19-23, 27.

责任编辑 王 浩