

网络信息体系信息流程有效低频路径挖掘方法

林文祥, 刘德生

引用本文

林文祥, 刘德生. [网络信息体系信息流程有效低频路径挖掘方法](#)[J]. 计算机科学, 2022, 49(11A): 211000001-6.

LIN Wen-xiang, LIU De-sheng. [Effective Low-frequency Path Mining Method for Information Flow of Networking Information-centric System of Systems](#) [J]. Computer Science, 2022, 49(11A): 211000001-6.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[一种基于禁忌搜索算法的流程挖掘方法](#)

Process Mining Approach Based on Tabu Search Algorithm

计算机科学, 2016, 43(4): 214-218. <https://doi.org/10.11896/j.issn.1002-137X.2016.04.044>

[流程增量挖掘中的模型更新方法](#)

计算机科学, 2009, 36(5): 154-157.

[基于Petri网的数据库日志分析方法研究](#)

Research on Database Log Based on Petri Nets

计算机科学, 2014, 41(6): 250-253. <https://doi.org/10.11896/j.issn.1002-137X.2014.06.049>

网络信息体系信息流程有效低频路径挖掘方法

林文祥 刘德生

航天工程大学复杂电子系统仿真重点实验室 北京 101416

(2385862812@qq.com)

摘要 随着信息技术和网络技术的迅猛发展及其在军事领域的广泛应用,网络信息体系应运而生。网络信息体系以信息为主导,主要的表现为其内含的信息活动流程。信息活动流程的合理性、高效性,直接影响信息在作战体系中的作战效能。采用流程挖掘技术从信息活动事件日志中发现信息活动流程模型,可为信息活动流程的建模、检验和增强提供有效支持。简单通过事件频率分析过滤日志中的噪声,容易导致有效低频路径丢失,降低挖掘的信息活动流程的准确性。结合军事信息活动的特殊性和信息传递的有效性特征,提出了一种基于结构聚合度的有效低频路径挖掘算法。仿真分析表明,该方法可成功分离日志噪声和有效低频路径,对挖掘真实有效的信息流程具有重要意义。

关键词: 流程挖掘; 网络信息体系; 信息流程; 有效低频; 结构聚合度

中图分类号 TP391

Effective Low-frequency Path Mining Method for Information Flow of Networking Information-centric System of Systems

LIN Wen-xiang and LIU De-sheng

Space Engineering University Science and Technology on Complex Electronic System Simulation Laboratory, Beijing 101416, China

Abstract With the rapid development of information technology and network technology and their widespread use in military field, networking information-centric system of systems comes into being. The networking information-centric system of systems is dominated by information, its main manifestation is the information activity process. The rationality and efficiency of the information activity process directly affect the operational effectiveness of information in the combat system. The use of process mining technology to discover information activity process models from information activity event logs can provide effective support for modelling, testing and enhancement of information activity processes. Simply filtering noise in logs through event frequency analysis can easily lead to the loss of valid low-frequency paths and reduce the accuracy of the mined information activity processes. Combining the special characteristics of military information activities and the effectiveness characteristics of information transfer, a structure aggregation degree based effective low frequency path mining algorithm is proposed. Simulation analysis shows that the method can successfully separate log noise and effective low frequency paths, which is important for mining real and effective information processes.

Keywords Process mining, Networking information-centric system of systems, Information process, Effective low frequency, Structure aggregation degree

1 引言

流程挖掘是一种可以从信息系统记录的事件日志数据中发现真实的结构化流程模型,并进行一致性检验和模型增强的技术^[1]。在全军抓建网络信息体系的大背景下,信息是主导,每一项信息活动都被记录在案。使用流程挖掘技术发现信息活动流程,对流程模型的分析 and 异常行为的查询和改进具有重要意义。而在真实的日志数据中,不可避免地会出现噪声数据,例如无序事件、异常行为或者记录错误^[2]。传统的启发式噪声处理思想是通过频率对噪声进行过滤,会对一些低频的有效路径产生误伤效应,降低了日志数据的准确性和完备性,影响流程挖掘的效果。

相关学者针对低频轨迹的处理进行了针对性研究。文献[3]提出了一种根据数据属性来分离低频有效路径和噪声的日志处理方法 DHM,以条件依赖度量为标准去除噪声;文献[4]采用 L* 生命周期和 DHM 作为工具,保留良好的不频繁关系;文献[5]结合频繁模式和交互行为特征来找出轨迹中的低频行为,提出了一种挖掘不频繁行为过程模型的优化方法;文献[6]提出了一种从流程模型中检索不频繁行为模式的新算法 WoMine-i,通过搜索提取低频的选择、并行、循环等结构;文献[7]提出了一种可扩展的框架,利用切操作过滤不频繁行为,能够处理非频繁行为和大规模事件日志;文献[8]提出使用整数线性规划的方法挖掘工作流网,并设计了对不频繁异常行为的过滤算法;文献[9]提出了一种基于 Petri 网

基金项目:国防科技重点实验室基础研究项目(DXZT-JC-ZZ-2018-002,DXZT-JC-ZZ-2017-001)

This work was supported by the Basic Research Project of National Defense Science and Technology Key Laboratory(DXZT-JC-ZZ-2018-002, DXZT-JC-ZZ-2017-001).

通信作者:刘德生(liudsnudt@126.com)

行为紧密度的低频有效路径挖掘方法,通过定义低频日志和模型的距离来判断行为的紧密程度,进而发现有效低频路径;文献[10]提出一种 WoSimp 算法,通过抽象日志中的非频繁行为来简化流程,以得到简单的流程模型;文献[11]和文献[12]基于机器学习的方法实现对流程挖掘中噪声的过滤。

上述方法都有着各自的优势,但也在一些方面存在着不足。文献[3-5]等对事件日志记录要求较高,需要得到行为的数据属性信息;文献[6-8]等需要已知完整的流程模型,且连同文献[10-12]均没有针对性地对有效低频路径进行挖掘;文献[9]根据向量的夹角余弦来判断向量相似性,当向量平行时,行为紧密度无法反映两个距离向量的相似程度。本文以网络信息体系下的联合作战军事信息活动流程为研究对象,在军事行动的特殊性和信息传递的有效性的双重约束下,提出了以结构聚合度为度量标准的有效低频路径挖掘算法,并通过实例验证了本文方法的可行性和有效性。

2 问题描述

战场形势瞬息万变,战争的不确定性必然会带来作战流程的不确定性,作战行动也必然会存在一些发生频率不高的流程路径,这些低频有效路径对战争局势有重要研究意义。比如某次作战活动中,为了提高指挥效率,上级指挥机构可能需要对基层单位进行直接指挥。应用流程挖掘技术挖掘信息流程,首先要区分低频有效路径和无效噪声路径。

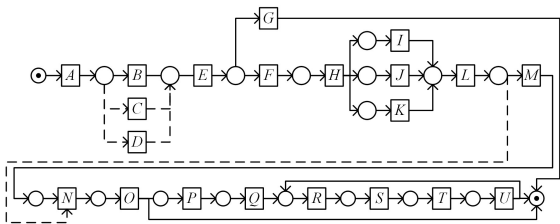


图1 弹道导弹战略预警拦截作战信息活动流程

Fig. 1 Information activity flow of ballistic missile strategic early warning and interception operations

不失一般性,本文以图1所展示的弹道导弹战略预警拦截作战信息活动流程为例,各节点信息活动如表1所列,整个案例表示的是一次发现敌情后的作战活动的信息流转情况。

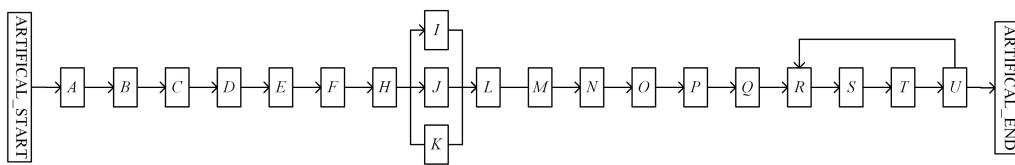
图1中,用虚线表示合理的低频行为,例如在正常情况下,高轨的红外卫星能率先探测到导弹发射的喷焰信息,但在部分特殊情况下,可由空基雷达或者天基雷达捕捉到发射症候;紧急情况下活动L略过M直接与N相连,缩短指挥周期,提高指挥效率。

表1 字母与信息活动对应表

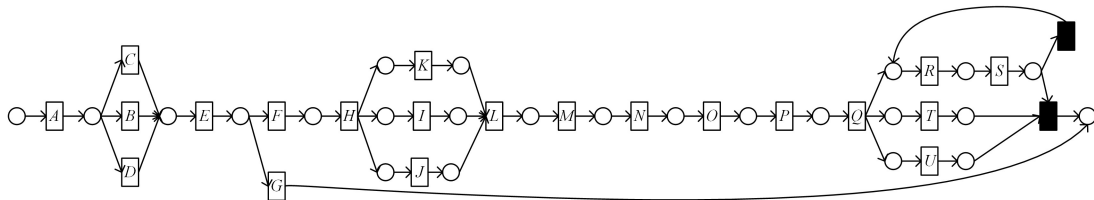
Table 1 Letters and their corresponding information activities

| 字母 | 信息活动名称 | 字母 | 信息活动名称 |
|----|-------------|----|---------|
| A | 作战单元配置 | L | 态势融合 |
| B | 高轨红外卫星探测到喷焰 | M | 态势理解 |
| C | 空基雷达捕捉到发射症候 | N | 目标识别 |
| D | 天基雷达捕捉到发射症候 | O | 威胁评估 |
| E | 特征匹配 | P | 拦截能力判断 |
| F | 发出告警信息 | Q | 作战优先级判定 |
| G | 虚警信号 | R | 制定拦截计划 |
| H | 高轨卫星引导 | S | 制导信息引导 |
| I | 低轨卫星跟踪 | T | 拦截弹状态跟踪 |
| J | 远程预警雷达跟踪 | U | 拦截效果评估 |
| K | 海基和舰载雷达跟踪 | | |

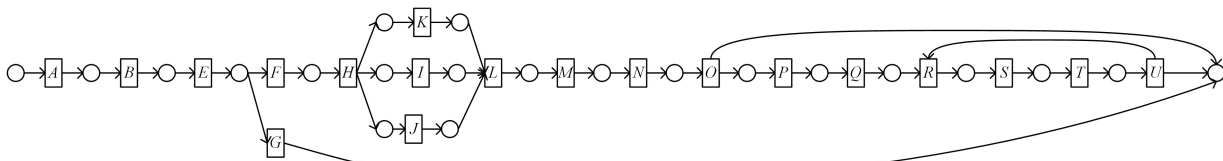
运用 CPN Tools 对上述信息流程进行仿真,在无噪声的条件下,产生包含 1000 条流程实例和 16745 个事件的事件日志数据,其中事件 C 和事件 D 的概率为 2%,事件 L 越级连接事件 N 的概率为 5%。分别使用 ProM 软件中对低频轨迹处理能力强的 HM 的插件、IMi 插件和 iDHM 插件对流程模型进行挖掘,均使用默认的过滤参数,得到流程模型如图 2 所示。与图 1 对比发现,HM 插件并不能挖掘出 ACE, ADE 和 LNO 路径,并且对于发生频率稍低的从 G 到结束和从 O 到结束的路径也不能挖掘出;IMi 不能挖掘出 LNO 路径和从 O 到结束的路径,且最后的 RSTU 相关活动的关系混乱;iDHM 插件在没有数据属性的条件下,不能挖掘出 ACE, ADE 和 LNO 路径。



(a) HM 插件挖掘的流程模型



(b) IMi 插件挖掘的流程模型



(c) iDHM 插件挖掘的流程模型

图2 各插件挖掘的流程模型

Fig 2 Process model of plug-in mining

由上述仿真结果可知,现有的流程挖掘插件内含的噪声处理算法,在没有噪声干扰的情况下,不能成功挖掘军事行动中的有效低频路径。本文提出一种基于结构聚合度的日志处理办法,实现对低频有效路径和无效噪声路径的区分,最后以有效低频路径优化流程模型,得到完整的信息流程。

3 基本概念

定义 1(事件日志) 设 E 是事件的有限集合,若 $l \in E^*$ 是一条实际执行的事件序列,则称 l 是一个流程实例,流程实例的集合 $L = \{l_1, l_2, l_3, \dots\}$ 被称为事件日志。其中, E^* 表示集合 E 的 Kleene 闭包。

在流程挖掘中,信息系统记录的事件日志是数据基础。例如,在事件集合 $E = \{e_1, e_2, e_3, e_4\}$ 中, $l_1 = \langle e_1, e_2, e_3, e_4 \rangle$ 和 $l_2 = \langle e_1, e_3, e_2, e_4 \rangle$ 是已经执行的事件序列,即分别是一个流程实例,而 $L = \{\langle e_1, e_2, e_3, e_4 \rangle^3, \langle e_1, e_3, e_2, e_4 \rangle^2, \langle e_1, e_2 \rangle\}$ 则是一个包含 6 个实例的事件日志,其中 $\langle e_1, e_2, e_3, e_4 \rangle$ 又被称为流程路径,是对轨迹相同的流程实例的总结,在右上角标注在日志中出现次数。

定义 2(流程实例的 k 阶继承关系) 设 $l_1 = \{e_1, e_2, e_3, e_4, e_5, e_6\}$ 是一个流程实例,则在该流程实例中,定义事件 e_1 和事件 e_2 的 k 阶继承关系为:存在整数 k ,使得事件 e_1 经过 $k-1$ 个事件可到达事件 e_2 ,记作 $e_1 \triangleright_k^l e_2$ 。

在一个流程实例中,由于循环结构等原因,可能存在多个 $k = (k_1, k_2, k_3, \dots)$ 值可使得 k 阶继承关系成立,定义流程实例最小 k 阶继承关系是当 k 值为 $k_{\min} = (k_1, k_2, k_3, \dots)_{\min}$ 时,记作 $a \triangleright_k^l b$ 。

例如,在流程实例 $l_1 = \langle e_1, e_2, e_3, e_4, e_5, e_6, e_4, e_7 \rangle$ 中,事件 e_3 和事件 e_4 的 k 阶继承关系是 $e_3 \triangleright_1^l e_4$ 和 $e_3 \triangleright_4^l e_4$,则该实例的最小 k 阶继承关系为 $e_3 \triangleright_1^l e_4$ 。

定义 3(日志的 K 阶继承关系) 设 $L = \{l_1, l_2, l_3, \dots\}$ 是一份事件日志,且事件 e_1 和事件 e_2 在各流程实例中的最小 k 阶继承关系是 $(e_1 \triangleright_{k_1}^l e_2, e_1 \triangleright_{k_2}^l e_2, e_1 \triangleright_{k_3}^l e_2, \dots)$,定义日志的 k 阶继承关系是存在 $K = (k_1, k_2, k_3, \dots)_{\min}$,使得该日志包含的所有流程实例的最小 k 阶继承关系中的 k 值均大于等于日志的 K 阶继承关系的 K 值,记作 $e_1 \triangleright_K^L e_2$ 。

定义 4(实例距离向量) 对于任意的流程执行实例 $l_1 = \langle e_1, e_2, \dots, e_{m+1} \rangle$, ($m \in \mathbb{N}^+$),定义该实例中相邻的两个事件的实例事件距离为 1,记作 $V_{l_1}(e_n, e_{n+1}) = 1 (1 \leq n \leq m)$,则该流程执行实例的实例距离向量为: $X = (1, 1, \dots, 1)_{1 \times m}$ 。

实例距离向量作为参考标准,主要是用于与日志距离向量进行对比,用单位向量表示,目的有两个:1)与日志距离向量的定义对应,当日志距离向量也为单位距离向量时,二者差异为零;2)便于后续的向量距离和结构聚合度的计算。

定义 5(日志距离向量^[9]) 设 $l_1 = \langle e_1, e_2, \dots, e_{m+1} \rangle$ ($m \in \mathbb{N}^+$) 是一条流程执行实例, $L = \{l_1, l_2, l_3, \dots\}$ 是一份给定的事件日志,定义 $\forall (e_n, e_{n+1}) \in l_1 (1 \leq n \leq m)$,且 $e_n \triangleright_k^L e_{n+1}$,其日志事件距离为 $V_L(e_n, e_{n+1}) = (e/2)^{K-1}$ (e 为自然底数),则该流程实例对应的日志距离向量为:

$$Y = \{V_L(e_1, e_2), V_L(e_2, e_3), \dots, V_L(e_m, e_{m+1})\}$$

$$= ((e/2)^{K_1-1}, (e/2)^{K_2-1}, \dots, (e/2)^{K_m-1})$$

在日志距离向量的定义中,考虑了越级连接的难度随着级数呈指数型增长,选择将 K 值作为指数,此外选择 $e/2$ 作为底数,是考虑越一级连接的事件日志距离不应过大。

定义 6(向量距离度量) 设流程实例 $l_1 = \langle e_1, e_2, \dots, e_{m+1} \rangle$ ($m \in \mathbb{N}^+$),事件日志 $L = \{l_1, l_2, l_3, \dots\}$,实例 l_1 的实例距离向量为 $X = (1, 1, \dots, 1)_{1 \times m}$,对应的日志距离向量为 $Y = ((e/2)^{K_1-1}, (e/2)^{K_2-1}, \dots, (e/2)^{K_m-1})$ 。采用 ffreys & Matusita 距离^[13]来度量两向量之间的差异,则流程实例 l_1 和事件日志 L 的向量距离度量定义为:

$$d(l_1, L) = \left[\sum_{i=1}^m (\sqrt{X_i} - \sqrt{Y_i})^2 \right]^{\frac{1}{2}}$$

$$= \left[\sum_{i=1}^m (1 - \sqrt{(e/2)^{K_i-1}})^2 \right]^{\frac{1}{2}}$$

其中, $m+1$ 是流程实例中的事件数, $(e/2)^{K_i-1}$ 是第 i 个事件和第 $i+1$ 事件的日志事件距离。向量距离 $d(l_1, L)$ 的大小反映了低频流程实例和低频日志的差异程度,距离越大,二者差异越大。

定义 7 将向量距离向量归一化,定义为结构聚合度,记作: $\xi(l_1, L) = 1/(1+d(l_1, L))$

事件之间的连接关系反映了事件执行的逻辑结构,这种结构的聚合度表征了某条路径与高频路径的吻合程度。对于低频路径而言,结构聚合度越高,说明该路径合理性越好,其取值在 0 到 1 之间,越接近 1 表示低频实例越接近高频日志,越有保留价值。

4 基于结构聚合度的有效低频日志挖掘算法

4.1 军事信息活动日志特性分析

面向网络信息体系背景下的信息流程挖掘,着眼于作战全局,目的性明确,注重研究整体的作战信息传递情况。由于军事家和战略家对战争的研究要具有前瞻性,面对复杂的战场环境和潜在的敌情要具有应对性,在战争设计阶段就要对有效低频行为有所考虑,这也是军事活动的特殊性,使得保留军事活动中的有效低频行为有迹可循,有科学依据和理论支撑。

本文将军事行动中的有效低频行为分为 3 种,分别是:

- (1) 备用单元的使用;
- (2) 备用路径的使用;
- (3) 合理的越级连接。

第(1)(2)种情况考虑的是作战行动中作战单元甚至部分作战路径功能丧失,需要引入备用单元和备用路径来保证行动的顺利进行,例如某次行动中,通信链路遭到敌方打击,需要引入备用的通信保障行动保持通信畅通;第(3)种情况是在上下级失联、某一级遭到敌方严重打击或者情势危急需要加快作战行动时出现,越级难度随着级数呈指数型上升。

4.2 算法相关设定

结合作战的实际情况和基于结构聚合度的有效低频路径挖掘方法的主要思想,设定如下:

- (1) 高频日志为有效行动的集合,可用作判断低频日志是否有效的依据。
- (2) 未按规定事件结尾的低频日志予以剔除,即认为该

日志未达到作战目的,属于异常行为。

(3)对连接备用单元和备用路径的低频行为予以保留,备用单元、备用路径与其他单元的日志的 K 阶继承关系的 K 值为 1。

(4)越级连接随着跨越级数的增加,难度呈现指数型增长,通过设定阈值去除不符合现实情况的越级连接情况,如要对越两级及以上的连接的路径予以剔除,其结构聚合度应该大于 $2/e$ 。

(5)当出现多次越一级的情况要慎重考虑其合理性,出现此类情况超过 3 次以上,可以认为其作战链路已经遭到严重破坏,丧失功能,不属于有效流程,可通过设定阈值去除噪声。

(6)考虑到符合越级连接和备用单元条件的流程路径也有可能是偶然的噪声,可以根据时序和频次关系对高于阈值的流程路径进行再次过滤,其依据就是在一场战役中,高频使用的流程路径被破坏,在短时间内抢修成功是不太现实的,备用路径一旦启用,在一段时间内,为保证军事行动的顺利执行,此路径将被重复使用。

4.3 基于结构聚合度的有效低频日志挖掘算法

基于上述分析和设定,根据军事行动特殊性和信息传递的有效性,采用 ffreys & Matsuta 距离定义结构聚合度,可将输入的事件日志数据去除噪声后,输出有效低频路径。

算法 1 基于结构聚合度的有效低频轨迹挖掘算法

输入:事件日志 $L = \{l_1, l_2, l_3, \dots, l_n\}$, 轨迹频率阈值 θ_1 , 结构聚合度阈值 θ_2

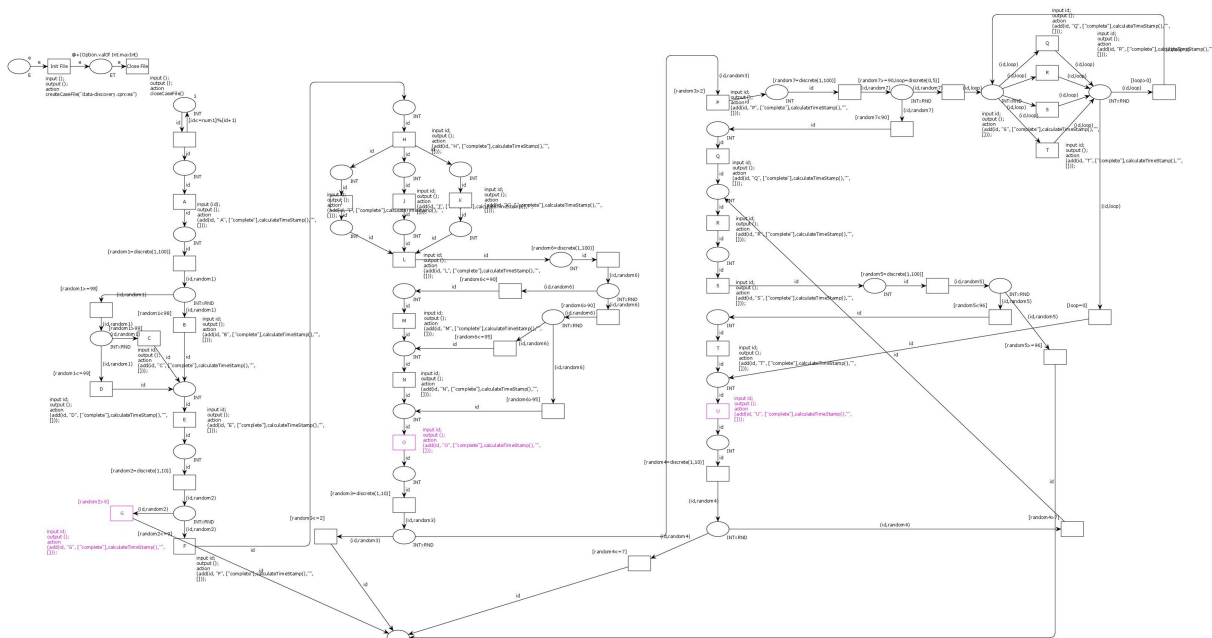


图 3 CPN Tools Petri 网仿真模型

Fig. 3 CPN Tools Petri network simulation model

运用 CPN Tools 进行 Petri 网仿真,在没有设定噪声的条件下,产生包含 1000 条流程实例和 16486 个事件的事件日志数据,原有设定不变,设置相关噪声路径,包括事件 L 越两级连接事件 O 的概率是 5%,事件 Q, R, S, T 会有 10% 的概率乱序连接,此外还设置了整体路径有 5% 的

输出:有效低频轨迹 $L' = \{\omega_1, \omega_2, \omega_3, \dots, \omega_m\}$

- Step1 将信息系统记录并生成的事件日志文件进行预处理,导入 ProM 操作平台,得到不同发生频率的流程路径;
- Step2 根据轨迹的频率阈值 θ_1 ,将事件日志分为有效的高频路径 $L_1 = \{\sigma_1, \sigma_2, \sigma_3, \dots, \sigma_q\}$ 和待处理的低频路径 $L_2 = \{\tau_1, \tau_2, \tau_3, \dots, \tau_p\}$;
- Step3 计算高频日志 $L_1 = \{\sigma_1, \sigma_2, \sigma_3, \dots, \sigma_q\}$ 事件之间的日志 K 阶继承关系,并把计算结果保留以待查用;
- Step4 判断 $L_2 = \{\tau_1, \tau_2, \tau_3, \dots, \tau_p\}$ 的日志完备性,删除不以规定事件结尾的低频路径,得流程路径 $L_3 = \{\tau_1, \tau_2, \tau_3, \dots, \tau_o\}$;
- Step5 遍历 $L_3 = \{\tau_1, \tau_2, \tau_3, \dots, \tau_o\}$,计算 τ_i 的实例距离向量;
- Step6 找出低频的流程实例中的相邻活动的日志 K 阶继承关系,计算日志距离向量,将其中有备用单元的事件日志距离定为 1,将找不到日志 K 阶继承关系两事件之间的日志距离定位 10,以便去除异常行为;
- Step7 根据 Step5 和 step6 得出的实例距离向量和日志距离向量计算结构聚合度,与阈值 θ_2 进行比较,大于阈值的流程实例予以保留,其余的当作噪声删除;
- Step8 判断是否已经遍历所有低频日志的流程路径,若未遍历完,则返回步骤 5;
- Step9 根据时序关系和频次进一步删除可能存在的噪声;
- Step10 输出有效的低频路径 $L' = \{\omega_1, \omega_2, \omega_3, \dots, \omega_m\}$ 。

5 实例验证

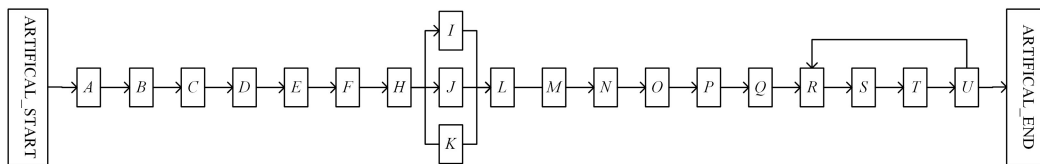
以弹道导弹战略预警拦击作战为例,其信息活动流程如图 1 所示,使用 CPN Tools 建立的仿真模型如图 3 所示。

概率以事件 S 异常结束。

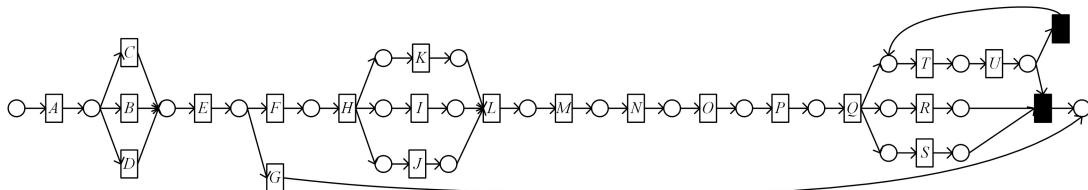
利用 ProM 软件的 HM 插件、IMi 插件和 iDHM 插件对含有有效低频路径和噪声的日志数据进行挖掘,参数均为默认值,挖掘结果如图 4 所示。对比发现:插件 HM 挖掘效果较未加噪声时有所下降,在随机噪声的干扰下,不能挖掘出从

U到R的再执行路径;插件IMi挖掘的模型在随机噪声的干扰下,关于R,S,T,U活动的执行情况有所变化,均未挖掘出

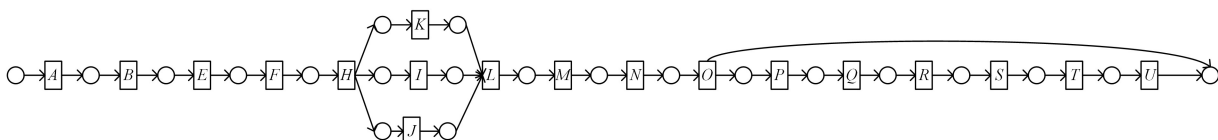
真实的路径;插件iDHM不能挖掘出从G到结束的路径,也不能挖掘出从U到R的再执行路径。



(a)HM 插件挖掘的流程模型



(b)IMi 插件挖掘的流程模型



(c)iDHM 插件挖掘的流程模型

图4 各插件挖掘的噪声条件下的流程模型

Fig. 4 Process model under noise conditions mined by various plug-ins

总体而言,3个插件在噪声的干扰下,无法有效区分噪声和有效低频路径。采用本文方法对同一日志数据进行处理,频率阈值为2%(低于该阈值的是低频路径),结构聚合度阈值为0.76(不保留3次以上的越一级连接和越二级以上的连接)。仿真结果表明:该方法可基本实现噪声和低频路径的分离(按表2的方式进行一一核对),部分结果如表2所列。

为了进一步验证有效低频路径挖掘的效果,使用挖掘的有效低频路径优化iDHM插件挖掘的流程模型,直至流程模型能够拟合所有的有效低频路径。其中,拟合表2展示的有效低频路径,即可实现完备流程的挖掘,如图5所示。

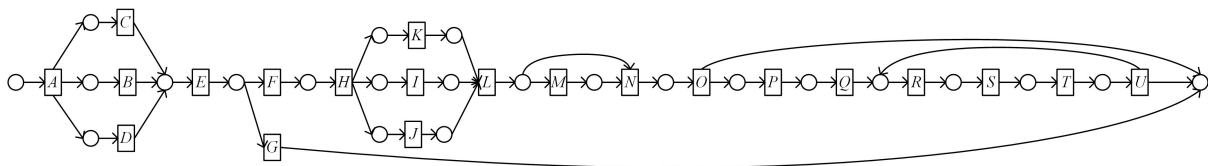


图5 优化后的完整模型

Fig. 5 Optimized complete model

在图4(c)iDHM插件挖掘的模型的基础上,根据表2中示例的有效低频路径,增加合理越一级连接路径LN,表示在紧急敌情迫使下,直接越级启动目标识别行动;增加备用单元D和E,表示通过空基雷达和天基雷达率先发现导弹发射信息;增加循环路径RSTU,表示拦截效果不好且时间充裕的前提下,对导弹实施二次打击;增加以G结尾的路径,表示识别判断为虚警信息后,解除告警信息。优化后的流程模型,与图1设计的理想模型保持了一致,展示了完整的信息流程,表明了基于结构聚合度的有效低频路径挖掘算法的有效性,对于后期的分析与研究有重要意义。

结束语 本文将流程挖掘技术应用于网络信息体系信息

表2 噪声和低频有效路径区分

Table 2 Difference between noise and low-frequency effective path

| 判断结果 | 流程路径 | 判断缘由 |
|--------|----------------------------|-------------|
| 有效低频路径 | ABEFHJKILMNOPQRSTU | 越一级连接 |
| | ABEFHJKILMNOPQRSTURSTURSTU | U到R的多次再执行路径 |
| | ADEFHKJILMNOPQRSTU | 备用单元D的使用 |
| | ADEG | 备用单元E的使用 |
| | | |
| 无效噪声路径 | ABEFHJKILMNOPRSSSSU | 乱序 |
| | ABEFHJKLO | 越两级连接 |
| | ADEFHJKILNOPQRS | 非规定事件结尾 |
| | | |

流程挖掘,提出了基于结构聚合度的有效低频日志处理算法。该算法实现了有效低频路径和无效噪声路径的成功分离,对挖掘真实且有分析意义的流程模型具有较好的参考价值。

在下一步的研究中,还需对算法进行进一步的验证。由于本文的真实数据难以获得,使用仿真数据进行算法验证,验证方式均一致,多次验证参考意义不大,在条件允许时需开展面向实战的信息流程挖掘研究,发现算法不足,增强算法的可用性。

参考文献

[1] VAN DER AALST W M P, WEIJTERS T, MARUSTER L.

- Workflow mining: discovering process models from event logs [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2004, 16(9):1128-1142.
- [2] SURIADI S, ANDREWS R, TER HOFSTEDE A, et al. Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs[J]. *Information Systems*, 2017(64):132-150.
- [3] FMANNHARDT, MD LEONI, REIJERS H A, et al. Data-Driven Process Discovery-Revealing Conditional Infrequent Behavior from Event Logs[C] // *International Conference on Advanced Information Systems Engineering*. Cham: Springer, 2017.
- [4] CHEN Q, LU Y, POONS K. An Algorithm to Preserve Infrequent Relations in Process Mining [C] // *Application to Lab Tests Ordering Process*. 2020.
- [5] WANG L L, FANG X W, ASARE E, et al. An Optimization Approach for Mining of Process Models with Infrequent Behaviors Integrating Data Flow and Control Flow[J]. *Scientific Programming*, 2021(1): 1-17.
- [6] CHAPELA-CAMPA D, MUCIENTES M, LAMAM. Discovering infrequent behavioral patterns in process models[C] // *International Conference on Business Process Management*. Cham: Springer, 2017: 324-340.
- [7] LEEMANS S J J, FAHLAND D, VAN DER AALST W M P. Discovering block-structured process models from event logs-a constructive approach[C] // *International conference on applications and theory of Petri nets and concurrency*. Cham: Springer, 2013: 311-329.
- [8] VAN ZELST S J, VAN DONGEN B F, VAN DER AALST W M P, et al. Discovering Relaxed Sound workflow nets using integer linear programming[J]. *Computing*, 2018, 100(5): 529-556.
- [9] GAO Y N, FANG X W, WANG L L. Business process configuration optimization analysis based on behavioral tightness of Petri nets[J]. *Computer Science*, 2017, 44(S1): 539-542.
- [10] CHAPELA-CAMPA D, MUCIENTES M, LAMA M. Simplification of complex process models by abstracting infrequent behaviour [C] // *International Conference on Service-Oriented Computing*. Cham: Springer, 2019: 415-430.
- [11] GOEDERTIER S, MARTENS D, VAN THIENEN J, et al. Robust process discovery with artificial negative events[J]. *Journal of Machine Learning Research*, 2009, 10: 1305-1340.
- [12] PONCE-DE-LEÓN H, CARMONA J, VANDEN BROUCKES K L M. Incorporating negative information in process discovery [C] // *International Conference on Business Process Management*. Cham: Springer, 2016: 126-143.
- [13] ZHANG Y, LIU Y D, JI Z. Vector similarity measurement method[J]. *Acoustics Technology*, 2009, 28(4): 532-536.



LIN Wen-xiang, born in 1996, postgraduate. His main research interests include system modelling and simulation, process mining.



LIU De-sheng, born in 1976, associate professor. His main research interests include system modelling and simulation and so on.