

基于流程切的过程模型挖掘方法

宋 健 方贤文 王丽丽

(安徽理工大学力学与光电物理学院 安徽 淮南 232001)

摘 要 在业务流程挖掘过程中,过程挖掘的目的是从事件日志中挖掘出满足人们需要的模型,以此来改善和优化过程模型。以往的研究都是从频繁日志中挖掘模型,将低频日志直接删除,该类方法使得挖掘的模型不完整,且在行为上会引发死锁或其他异常情况。文中提出基于流程切的过程模型挖掘方法,该方法从事件日志中挖掘过程模型,对事件日志采用流程切的形式进行分割,不仅考虑到频繁行为,还考虑了低频模式下的行为;尤其针对异常的环状结构会引起流程图的边缘结构发生异常的问题,流程切可以很好地进行处理。利用这种方法得到的模型比较全面完善,能够提高有效性和精确度。利用评价指标对构建的模型进行优化,从而得出最优模型。最后,通过具体事例验证了所提方法的有效性。

关键词 流程切,过程挖掘,低频模式,事件日志,Petri 网

中图法分类号 TP391.9 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.07.048

Process Model Mining Method Based on Process Cut

SONG Jian FANG Xian-wen WANG Li-li

(College of Mechanics and Optoelectronics Physics, Anhui University of Science and Technology, Huainan, Anhui 232001, China)

Abstract The purpose of process mining is to mine the model that meets people's needs from the event log in the mining process of business process, so as to improve and optimize the process model. In previous studies, models were mined from frequent log, and low-frequency log was deleted directly, which makes the mined model incomplete and may cause a deadlock or other abnormalities. This paper proposed a process model mining method based on process cutting. The method mines the process model from the event log and segments the event log in the form of a process cutting, not only taking into account the frequent behaviors, but also the behaviors in the low-frequency mode. In particular, aiming at the problem that the abnormal circular structure will cause abnormality for edge structure of the flow chart, the process cutting can handle it well. The model obtained by proposed method is more comprehensive and perfect, and the validity and accuracy of the model can be improved. The evaluation index was used to optimize the constructed model and the optimal model was obtained. Finally, the effectiveness of the method was verified by concrete examples.

Keywords Process cut, Process mining, Low frequency mode, Event log, Petri net

1 引言

过程挖掘的目的是改进业务流程。在以往挖掘过程中,主要是对频繁行为模式进行挖掘,并对频繁行为事件日志进行监视,目的是从事件日志中提取相关信息,研究相关实例所要发生的流程关系。随着业务流程的发展,从事件日志中发现不频繁行为也越来越重要,因为它可以揭示流程中不常见的错误,能够发现行为异常或者行为偏差,促使人们及时改正流程,降低模型的复杂性。低频模式挖掘对业务流程的管理也至关重要,能够使流程模型变得更加完善,更加符合企业管理者的需求,有利于提高流程管理的生产效率。因此,从事件

日志中挖掘低频模式的行为已成为当下的热门研究课题之一。

在业务流程挖掘过程中,流程挖掘的目的是从事件日志中挖掘出有用的信息。将源模型进行还原,保证流程模型与实际生活中的业务流程相一致,并对实际流程进行监视和改进,使模型的完备性得到提高。文献[1-2]提出通过事件序列的形式观察过程执行的模型,并开发发现 workflow 模型的技术,该技术的起点是一个“工作流日志”,包含正在执行的相关工作流程过程的信息;同时提出一种新的算法从这些日志中提取过程模型,并用 Petri 网表示。文献[3-4]针对不可见任务,提出产生不可见任务所需的条件,并对不可见任务给出了相

到稿日期:2018-06-01 返修日期:2018-07-19 本文受国家自然科学基金资助项目(61572035,61272153,61402011),安徽省自然科学基金资助项目(1508085MF111),安徽省高校自然科学基金资助项目(KJ2014A067,KJ2016A208)资助。
宋 健(1991—),男,硕士生,主要研究方向为 Petri 网,E-mail:1529139842@qq.com;方贤文(1975—),男,博士,教授,主要研究方向为 Petri 网和可信软件,E-mail:280060673@qq.com(通信作者);王丽丽(1982—),女,副教授,主要研究方向为业务流程管理。

关检测方法。文献[5]提出在发现过程模型的过程中,通过搜索并删除非频繁行为(偏差或异常迹)来降低模型的复杂性;同时还提出了一种自动删除事件日志中非频繁行为的技术,有些低频率序列的日志属于噪音序列,另一些低频率序列的日志则属于有效低频率序列范畴。对于有效低频率序列,若直接对其进行过滤处理,则会导致模型的不完整性并降低模型的质量(结构性、行为性、精确度)。因此,从日志序列中提取非频繁行为至关重要,能极大地提升模型的质量。文献[6]提出一种启发式挖掘算法,该算法在构建过程模型时考虑了事件日志的发生频率。文献[7]提出一种新的建模方法,即基于事件日志的流程挖掘方法,根据事件日志记录的活动执行顺序构建一个标准的业务流程模型。文献[8-9]提出了一些过程挖掘方法,除常见的 α 算法外,还包括一种基于区域的过程挖掘算法。文献[10]通过迹的活动片段得到子模块的流模型,并对挖掘出的子流程模型进行准确度的判断,从而确保流程模型的精确性。文献[11-12]提出了一种新的算法——WoMine-i,该算法能够从过程模型中检索出所有的低频行为模式,同时可以在选择、并行、循环的序列中得以实现。已有的低频挖掘方法过于粗糙、冗余,在处理低频事件日志时,很多都将低频率序列当作噪音或者行为异常直接过滤处理,很少考虑到日志的有效性。虽然有些事件日志发生的频率不是太高,但是在业务流程环节中,它是有意义并且有利的,因此我们应该保留这类事件日志。

本文以行为轮廓为基础,基于流程切的方法从事件日志中挖掘有效的低频率序列日志。首先,根据给定的事件日志序列将高频率日志优先筛选出来,并将其作为输入,根据行为轮廓理论,构建出行为轮廓矩阵关系表,基于行为轮廓关系表便可以构建初始 Petri 模型。然后,利用流程切对日志进行分割处理,包括严格序分割、排他序分割、交叉序分割和循环序分割。在这些分割的模块中,将不合理的低频率序列模块直接删除,保留未删除的低频日志,进而构建低频率序列子模块并将其添加到源模型中,以对源模型进行补充优化。最后,通过计算比较源模型与新模型的一致性以及行为适当性,说明所挖掘的低频率序列为有效的低频率序列,从而达到优化和完善模型的目的,使模型变得更加完整。

2 研究动机

本文利用流程切对日志序列进行挖掘,将序列日志按交叉序、排他序、循环序和严格序进行分类处理,并做出日志序列的流程切。利用虚线将流程切进行切割,通过各个日志之间的行为轮廓关系建立初始模型。流程切切割日志序列方法如下,给定序列日志 $L: [\langle b, g, i \rangle^{50}, \langle b, g, h, g, i \rangle^{60}, \langle a, c, d, e, f, g, i \rangle^{100}, \langle a, c, e, d, f, g, h, g, i \rangle^{80}, \langle a, c, e, f, g, i \rangle^{80}, \langle a, c, e, f, g, h, g, i \rangle^{120}, \langle a, c, d, e, f, g, i \rangle^{200}, \langle a, c, e, d, f, g, i \rangle^{150}]$, 在流程切中,一个节点表示一个活动,若活动 a 在日志 L 中直接跟随 b ,则从节点 a 到节点 b 用弧线(直线)表示该活动。流程切中有 4 种操作符: $\rightarrow$ 表示顺序组合操作符, $\wedge$ 表示交叉操作符, $\times$ 表示排他操作符, $\otimes$ 表示循环操作符。日志的流程切如图 1(a)所示,利用 $\rightarrow$ 操作符将日志按照 $\{a, b, c, d, e, f\}, \{g, h, i\}$ 从左至右分割,得到左分支日志 $L_1 = [\langle b \rangle^{110}, \langle a, c,$

$d, e, f \rangle^{300}, \langle a, c, e, d, f \rangle^{230}, \langle a, c, e, f \rangle^{200}]$ 以及右分支日志 $L_2 = [\langle g, i \rangle^{580}, \langle g, h, g, i \rangle^{260}]$,日志 L_1, L_2 得到的流程切见图 1(b)和图 1(c)。将日志 L_2 利用 $\otimes$ 分割,得到日志 $L_3 = [\{g, h\}^{260}, \{h, g\}^{260}]$,日志 L_3 的流程切见图 1(d)。流程切将日志分成若干子日志序列后,利用各子日志序列之间的行为轮廓关系构建初始模型 M_0 ,挖掘出的 Petri 网模型如图 1(e)所示。

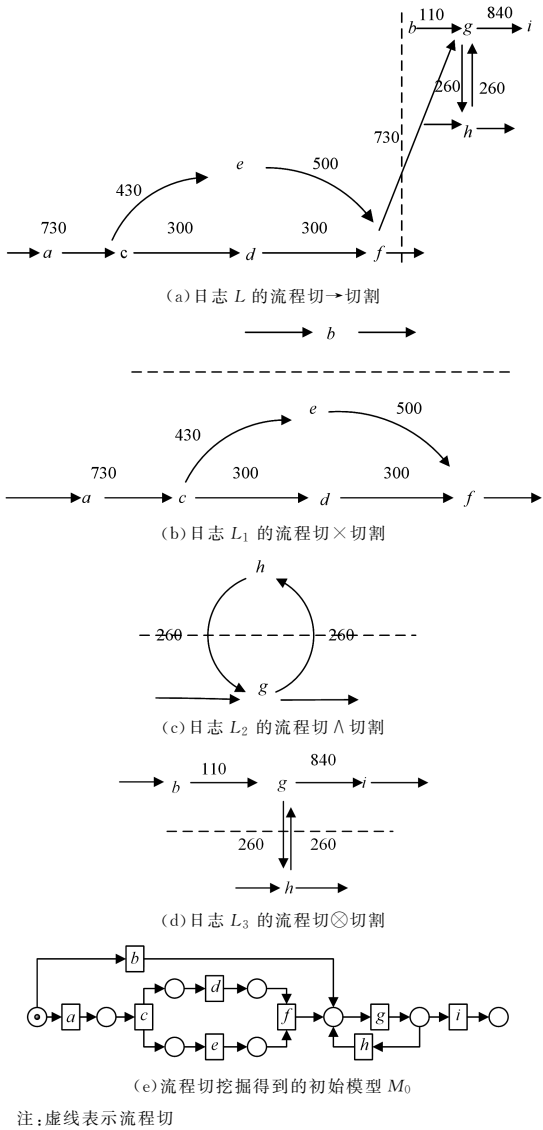


图 1 4 种模式的流程切切割图

Fig. 1 Process cutting graph of four modes

基于以上提出的 4 种操作技术,流程切能够对日志进行很好的过滤处理。基于日志来分析环状结构对于研究者来说一直是一个棘手的问题,因为环状结构分析比较复杂,而且容易引发其他结构上的问题。但针对环状结构,流程切能够很好地进行处理,且方法简单、快速、精确。日志 $L_1 = [\langle a, c, d, e, b \rangle, \langle a, b, a, e, d, c \rangle, \langle a, e, c, b, d \rangle, \langle a, b, d, c, e \rangle]$,从日志中可以发现第二条迹中存在非频繁行为 a 。利用流程切的方法做出日志 L_1 的直接流关系的切割(见图 2(a)),同时做出基于间接流关系的流程切割图(见图 2(b))。从流程切图中可以发现,基于直接流关系的传递闭包切割图中,边缘 $\langle a, b \rangle$ 表示序列活动 b 紧随在活动 a 后发生。在图 2(a)的直接流关系

切割图中,活动 b 的所有输出边缘的弧值都为 1。在图 2(b)的间接关系切割图中,除了非频繁活动 $\langle b, a \rangle$ 的弧值没有改变外,其他的输出边缘流关系的弧值都被放大,因此可以将边缘活动 $\langle b, a \rangle$ 过滤。

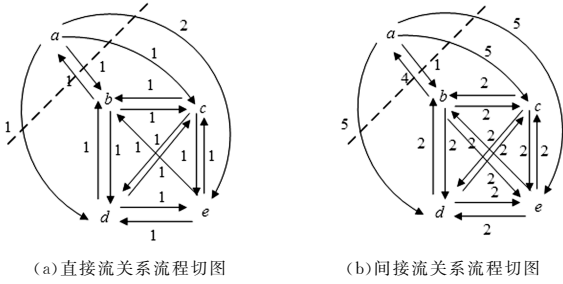


图 2 基于流程切的环状结构过滤切割图

Fig. 2 Filter cutting diagram of circular structure based on process cutting

3 基本概念

定义 1(流程模型 Petri 网)^[13] 一个流程模型 Petri 网 $PN=(P, T; F, C)$ 是一个四元组,满足以下条件:

- (1) $P \cup T \neq \emptyset$;
- (2) $P \cap T = \emptyset$;
- (3) $F \subseteq (P \times T) \cup (T \times P)$;
- (4) $dom(F) = \{x \in P \cup T \mid \exists y \in P \cup T: (x, y) \in F\}$;
- (5) $cod(F) = \{x \in P \cup T \mid \exists y \in P \cup T: (y, x) \in F\}$ 。

可见,网 $PN=(P, T; F, C)$ 的基本元素集合为 P 和 T ,在 Petri 网中,一般分别使用圆圈和方框来表示。

定义 2(行为轮廓)^[14] $PN=(P, T, F)$ 是一个 Petri 网,任意的变迁对 $(x, y) \in (T \times T)$ 满足下列关系之一:

- (1) 严格序关系 $\rightarrow$, 当且仅当 $x > y, y \not> x$;
- (2) 排他序关系 $+$, 当且仅当 $x \not> y, y \not> x$;
- (3) 交叉序关系 $\parallel$, 当且仅当 $x > y, y > x$ 。

则以上几种行为关系构成 Petri 网的行为轮廓。其中, $\leftarrow$ 表示逆严格序关系。

定义 3(事件日志)^[1] T 是任务集, $\sigma \in T^*$ 是一个执行迹, $L \in P(T^*)$ 是一个事件日志。 $P(T^*)$ 是 T^* 的幂集, $L \subseteq T^*$ 。

定义 4(低频模式)^[15] 设 L 是过程日志的迹集, 一个简单模式 SP 的频率为 $\frac{freq(SP) = |\{\tau \in L; SP \vdash \tau\}|}{|L|}$, 而模式

P 的频率为简单模式的最大频率, 即: $freq(P) = \max freq(sp), \forall sp \in P$ 。给定一个频率阈值 $\sigma \in R; 0 < \sigma \leq 1$, 一个模式 P 是一个低频模式, 当且仅当 $freq(P) < \sigma$ 。

定义 5(流程树)^[12] 设 $L_i (i = 1, 2, \dots, n)$ 为流程模型 $CP=(S, T, F, c)$ 对应的事件日志, $S(T)$ 为变迁 T 的迹, $L \in S(T), c \subseteq S \cup T$, 如果变迁对 $\forall x, y \in c: ((x, y \notin F^+) \wedge (y, x \notin F^+))$, 其中 F^+ 为流关系 F 的传递闭包。对于任意一个活动对 $(x, y \in L_i \times L_j), \forall \sigma = t_1 t_2 \dots t_n$, 其中 $i, j \in \{1, 2, \dots, n\}, i < j$, 其满足下列关系中的一种:

- (1) 弱序关系 $x >> y$, 当且仅当 $\exists \sigma \in L \Rightarrow t_i = x, t_j = y$;

(2) 因果关系 $x \rightarrow y$, 当且仅当 $x >> y, y \neq x$;

(3) 无关系 $x << y$, 当且仅当 $x \not>> y, y \not>> x$;

(4) 平行交叉关系 $x \parallel y$, 当且仅当 $x >> y, y >> x$;

(5) 互斥关系 $x \rightarrow y$, 当且仅当 $\forall \sigma \in L, \forall x, y \in T \Rightarrow (x \in \sigma, y \notin \sigma) \vee (y \in \sigma, x \notin \sigma)$ 。

定义 6(适合度)^[14] 设 $L_P = n_1, n_2, \dots, n_m$ 是流程模型 $P=(A, C, T, F)$ 的一组日志。集合 $SR \subseteq (A_L \times A_L)$ 包含活动变迁对 (x, y) , 日志 L_P 的行为关系映射到流程模型 P 中, $SR(R_P, R_L)$ 满足 $R_P \in \{\rightarrow, \rightarrow^{-1}\} \wedge R_L = \times$ 或 $R_P = R_L$ 或 $R_P = \wedge$, 则日志 L_P 在流程模型 P 中的重放适合度 $\alpha L_P = |S_R| \setminus (|A_L \times A_L|)$ 。同时还要考虑流程模型的行为适当性, 根据

公式 $\alpha_B = 1 - \frac{\sum_{i=1}^k n_i (x_i - 1)}{(m-1) \sum_{i=1}^k n_i}$ 进行验证。其中, k 为日志中不同的

轨迹数, n 为日志轨迹中所含的数目, x 代表日志轨迹重放时变迁的平均数目, m 为模型中可见任务的个数。

定义 7(一致性度量)^[16] 设 $(N_1, [i_1])$ 和 $(N_2, [i_2])$ 是两个工作流系统, 且关于 $\sim$ 匹配。 L_1, L_2 分别是变迁集 T_1, T_2 上的多重变迁集 (有 $T_1 \subseteq L_1, T_2 \subseteq L_2$), 且 $CL_1 \subseteq (\tilde{L}_1 \times \tilde{L}_1), CL_2 \subseteq (\tilde{L}_2 \times \tilde{L}_2)$ 分别是工作流的一致性度量的多重变迁集, 则 $\sim$ 的基于多重变迁集的行为轮廓的一致性度量公式为:

$$MBP_{\tilde{L}} = \frac{\omega_1 |\tilde{L}_1 \times \tilde{L}_1| + \omega_2 |\tilde{L}_2 \times \tilde{L}_2|}{|\tilde{L}_1 \times \tilde{L}_1| + |\tilde{L}_2 \times \tilde{L}_2|}$$

其中:

$$\omega_1 = \frac{\sum_{(t_x, t_y) \in CL_1} (\tilde{\delta}(t_x) + \delta(t_y))}{\sum_{(t_x, t_y) \in (\tilde{L}_1 \times \tilde{L}_1)} (\tilde{\delta}(t_x) + \delta(t_y))}$$

$$\omega_2 = \frac{\sum_{(t_i, t_j) \in CL_2} (\tilde{\delta}(t_i) + \delta(t_j))}{\sum_{(t_i, t_j) \in (\tilde{L}_2 \times \tilde{L}_2)} (\tilde{\delta}(t_i) + \delta(t_j))}$$

$$\tilde{\delta}(t) = \frac{\sum_{(t_x, t_y) \in \sim \wedge (t=t_x \vee t=t_y)} \delta(t_x, t_y)}{|\{(t_x, t_s) \in \sim \mid t=t_x \vee t=t_s\}|}$$

4 流程切挖掘低频率序列日志的方法

过程挖掘中, 大部分方法都是对频繁模式进行研究, 目的是增强和改进业务流程。通过对频繁模式的研究, 可以检测业务流程存在的异常和偏差, 以达到监视和优化的目的。对于低频模式的行为, 当前很少有学者进行相关研究。低频率序列日志的发生率较低, 且大部分都是无效的, 所以很多研究都将其作为噪音过滤掉。但在实际的案例研究中发现, 不是所有的低频行为模式都是无效的。因此, 从这些低频率序列中筛选出有效的低频模式至关重要, 这些低频对完善业务流程以及增强模型的稳定性非常关键。本文基于流程切的方法, 在事件日志中对日志序列进行挖掘, 将低频日志模式筛选出来并重放到初始模型中, 计算模型的一致性度量值、模型的适合度、行为适当性等标准, 直到挖掘出符合标准并且完善的业务流程模型。具体的挖掘算法如算法 1 所示。

算法 1 找出符合流程模型的低频率序列的算法

输入: 事件全部的日志序列 L , 流程切弧值 r , 弧值频率 ξ ;

输出: 符合流程模型的低频模式序列

- 步骤 1 为防止重复操作,首先对所有的日志进行预处理,将相同的日志进行合并操作。
- 步骤 2 利用流程切将所有的日志按照 4 种模式进行切割操作,将不合理的日志序列切割。
- 步骤 3 完成流程切处理后,将日志按发生频率的大小重新排列,将发生频率大于 2%(该频率发生次数与总日志数的比值)的日志序列筛选出来,分析各个变迁对 t_1, t_2 (其中 $t_1, t_2 \in L$) 的行为轮廓关系,构建活动关系表。
- 步骤 4 根据行为轮廓表以及所有的事件日志系列,做出直接流关系的流程切图和间接流关系的流程切图,直接流关系的弧值设为 r_i ,间接流关系的弧值设为 r_j ,若 $r_j > r_i$,则保留该部分的日志序列,若 $r_i \leq r_j$,则直接将其过滤删除。该步骤能够过滤掉不合理的序列日志和不合理的环状结构。
- 步骤 5 将不合理的低频序列日志删除后,得到了符合条件的事件日志。重新将这些符合要求的日志按照频率进行排列。
- 步骤 6 在步骤 3 中,构建活动关系表时并未考虑日志发生频率小于 2% 的日志,如 $L_i, L_j, \dots$,根据低频模式的定义,计算日志的频率 $\text{freq}(i) (i=1,2,3,\dots)$ 。若事件日志序列 $\text{freq}(j) \geq \xi_r$,则保留该日志;若 $\text{freq}(j) < \xi_r$,则将该日志序列过滤掉。
- 步骤 7 迭代计算步骤 6,对所有发生频率小于 2% 的日志计算完成后,得到一组新的过滤日志,当前的这些日志就是我们所需要的低频模式下的序列日志。
- 步骤 8 完成步骤 7 的计算,删除低频模式的冗余信息,输出流程模型的低频模式序列,算法结束。

基于算法 1 的低频模式序列挖掘,得到符合要求的低频模式及序列。利用流程切的方法对日志进行分类切割处理,将事件日志的活动关系用流程切的关系流图表示出来,找出变化不一致的区域,并利用低频率序列日志进行补充。再对模型的行为适当性以及模式的一致性度量值进行计算验证,从而找出最优的过程模型。基于流程切的过程模型挖掘算法如算法 2 所示。

算法 2 基于流程切的过程模型挖掘算法

输入:处理过的事件日志 L, 阈值 t_f

输出:Petri 网模型

- 步骤 1 将算法 1 得到的日志按发生频数从小到大依次排列。利用流程切的方法将日志序列进行分类,并且对日志序列进行切处理,同时建立初始 Petri 网流程模型 M_0 。
- 步骤 2 对日志进行流程切处理后,将事件日志的活动对关系用流程切关系图表示出来,并判断初始模型序列关系与流程切图关系是否相匹配,若匹配则保留,否则将不完备的事件日志轨迹删除。
- 步骤 3 计算初始流程模型的适合度 $\epsilon_{LP}(M_0)$ 。若 $\epsilon_{LP}(M_0) \geq t_f$,则模型合理,执行步骤 4;若 $\epsilon_{LP}(M_0) < t_f$,则增加配置信息,调整流程模型,直至流程模型满足 $\epsilon_{LP}(M_0) \geq t_f$ 。
- 步骤 4 根据定义 6 计算日志与模型的行为适当性 α_B 的值,若 $\alpha_B(M_0) < 0.9$,则说明日志与模型的行为适当性偏低。将算法 1 挖掘得到的低频模式序列重放到流程中以优化模型,得到优化后的模型 M_1 ,令 $M_0 = M_1$,再返回步骤 3。
- 步骤 5 经过步骤 4 优化得到模型 M_1 ,利用增量日志的弱序关系,将模型 M_1 与初始模型 M_0 的行为轮廓关系进行对比,若行为一致,则执行步骤 6,否则在不一致的地方添加配置变迁,并且删除不一致的路径以调整模型,从而得到模型 M_1 ,

- $M_2, \dots, M_i, M_j, \dots$ 。
- 步骤 6 根据定义 7 计算模型 $MBP_L[M_j, M_0]$ 的行为轮廓一致性度量值,若 $MBP_L[M_j, M_0] \geq MBP_L[M_i, M_0]$,则将 M_j 作为新的模型;若 $MBP_L[M_j, M_0] < MBP_L[M_i, M_0]$,则将 M_i 作为新的模型。
- 步骤 7 设置模型适合度 α_{LP} 的参数 W_i 和行为适当性 α_B 的参数 W_k 的值,计算 $Q(M) = \frac{W_i \alpha_{LP}(M) + W_k \alpha_B(M)}{W_i + W_k}$,通过设置不同权重 W_i, W_k 的值,使得 $Q(M_1) > Q(M_0)$ 。
- 步骤 8 所有日志重放结束,直至 $\alpha_{LP}(M_0) < \alpha_{LP}(M_1), \alpha_B(M_0) < \alpha_B(M_1)$,则 M_1 为所要挖掘的流程模型。

5 案例分析

本节通过一个简单的实例,即基于事件日志挖掘导航系统的业务流程模型来验证算法的可行性。记录的事件日志分别用以下大写字母表示:A:打开导航;B:输入目的地;C:提供路线方案;D:线路一;E:线路二;F:线路三;G:收费站 c_1 ;H:收费站 b_1 ;I:收费站 a_1 ;J(L,N,P):小道;K:收费站 a_2 ;M:收费站 b_2 ;O:收费站 c_2 ;Q:收费站 b_3 ;R:收费站 a_3 ;S:收费站 a_4 ;T:到达目的地。日志的执行轨迹以及实例数如表 1 所列。

表 1 事件日志
Table 1 Event logs

日志 L_1		日志 L_2	
日志轨迹	实例数	日志轨迹	实例数
ABCDIJST	680	ABCFGOT	1656
ABCDIKRST	1505	ABCFNMQT	356
ABCEHMQT	1355	ABCEHMQT	308
ABCFGPT	450	ABCDIJST	452
ABD	8	ACD	12
日志 L_3		日志 L_4	
日志轨迹	实例数	日志轨迹	实例数
ABCDIKRST	1346	ABCEHMQT	198
ABCDILQT	613	ABCFGOT	1236
ABCEHMQT	1152	ABCDILQT	548
ABCFGPT	396	ABCEPT	98
BCDT	3	ABCFNMQT	298

根据算法 1 的执行步骤,首先将相同的日志事件轨迹进行合并,并且将事件日志序列按照实例数从多到少进行排列,结果如下: $\{\langle ABCFGOT \rangle^{2892}, \langle ABCDIKRST \rangle^{2851}, \langle ABCEHMQT \rangle^{2507}, \langle ABCDILQT \rangle^{1161}, \langle ABCDIJST \rangle^{1132}, \langle ABCFGPT \rangle^{846}, \langle ABCFNMQT \rangle^{654}, \langle ABCEHMQT \rangle^{506}, \langle ABCEPT \rangle^{98}, \langle ACD \rangle^{12}, \langle ABD \rangle^8, \langle BCDT \rangle^3\}$ 。为了更好地观察日志序列的大小,绘制出事件日志的分布直方图,如图 3 所示。

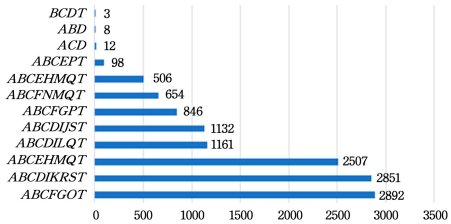


图 3 事件日志条形图

Fig. 3 Event log bar graph

选取发生频率高的事件日志 $\{\langle ABCFGOT \rangle^{2892}, \langle ABCDIKRST \rangle^{2851}, \langle ABCEHMQT \rangle^{2507}, \langle ABCDILQT \rangle^{1161}, \langle ABCDIJST \rangle^{1132}, \langle ABCFGPT \rangle^{846}, \langle ABCFNMQT \rangle^{654}, \langle ABCEH-$

$MQT\}^{506}$,根据定义 2,计算改进后的事件日志行为轮廓关系,并建立行为轮廓关系表,具体如表 2 所列。

表 2 行为轮廓关系表

Table 2 Behavior profile relationship table																				
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
A	+	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→
B		+	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→
C			+	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→
D				+	+	+	+	+	→	→	→	+	+	+	+	→	→	→	→	→
E					+	+	+	→	+	+	+	+	→	+	+	+	→	+	+	→
F						+	→	+	+	+	+	+	→	→	→	→	→	+	+	→
G							+	+	+	+	+	+	→	+	→	→	→	+	+	→
H								+	+	+	+	→	+	+	+	→	→	+	+	→
I									+	→	→	→	+	+	+	→	→	→	→	→
J										+	+	+	+	+	+	+	+	+	→	→
K											+	+	+	+	+	+	→	→	→	→
L												+	+	+	+	→	→	→	→	→
M													+	+	+	→	→	→	→	→
N														→	+	+	→	→	→	→
O																+	+	+	→	→
P																	+	+	→	→
Q																		+	→	→
R																			+	→
S																				+
T																				

同时根据所选取的高频日志序列,计算出这些日志的活

动关系表,具体如表 3 所列。

表 3 活动关系表

Table 3 Activity relationship table																				
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
A	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	3	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
K	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

根据行为轮廓关系表 2 和活动关系表 3,利用第 2 节流程切的方法,建立初始模型 M_0 ,得到的模型如图 4 所示。

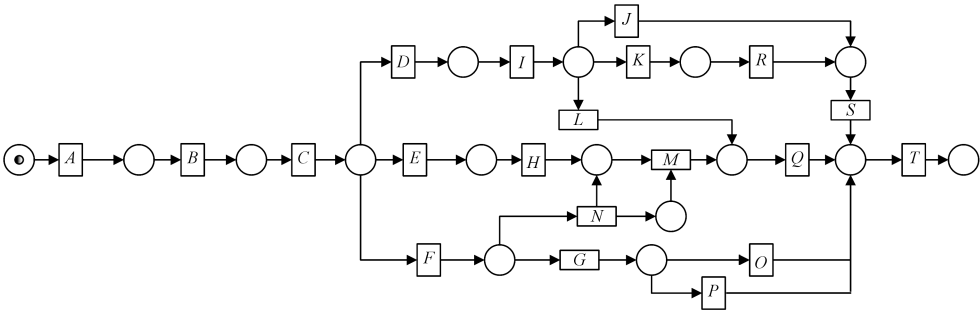


图 4 初始模型业务流程 Petri 网 M_0

Fig. 4 Initial model business process Petri net M_0

初始模型 M_0 是由高频率列日志得到的,其中并未考虑所有的低频日志,因此得到的模型 M_0 并不是最完善的模型。根据实际需求,将某些有效的低频率列日志考虑到模型中。针对模型的实际意义,可以将 $\langle ABD \rangle, \langle ACD \rangle, \langle BCDT \rangle$ 日志删除,因为其不满足实际意义,所以将其视为噪音进行处理。而日志 $\langle ABCEPT \rangle$ 则需要进一步判断,不能直接将其删除。基于定义 4,计算日志 $\langle ABCEPT \rangle$ 的低频模式值 $freq =$

$0.12 > 0.1$,根据定义 6 模型的适合度 $aL_P = |SR|/(A_L \times A_L) = \frac{6 \times 6}{6 \times 6} = 1$,所以日志重放到初始模型中的适合度是满足的。

根据行为轮廓关系表 2 以及日志活动关系表 3,通过添加增量变迁 X 进行配置,挖掘到增量模块 S_0 ,见图 5 中的虚线部分,同时将得到的增量模块嵌入到初始模型 M_0 中,对初始模型进行补充完善,得到的优化模型 M_1 如图 5 所示。

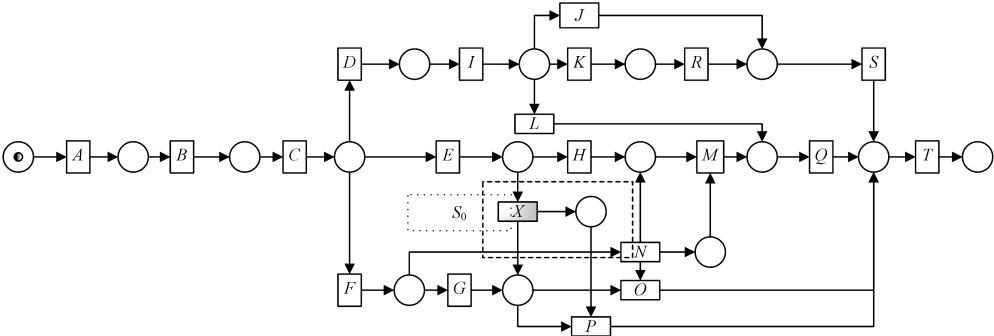


图 5 目标模型业务流程 Petri 网 M_1

Fig. 5 Target model business process Petri net M_1

根据定义 6 提出的行为适当性计算公式,对所得到的初始模型 M_0 和目标模型 M_1 分别进行计算验证,计算结果如下:

$$\alpha_B(M_0)=1-\frac{\sum_{i=1}^kn_i(x_i-1)}{(m-1)\sum_{i=1}^kn_i}$$
$$=1-(\frac{7*6+9*(\frac{7+9}{2})+8*(\frac{7+9+8}{3})+8*(\frac{7+9+8+8}{4})+8*(\frac{7+9+8+8+8}{5})}{19*(7+9+8+8+8+8+7+8+8)}+\frac{7*(\frac{7+9+8+8+8+7}{6})+8*(\frac{7+9+8+8+8+7+8}{7})+8*(\frac{7+9+8+8+8+7+8+8}{8})}{19*(7+9+8+8+8+8+7+8+8)})$$
$$=1-0.4067=0.5933$$
$$\alpha_B(M_1)=1-\frac{\sum_{i=1}^kn_i(x_i-1)}{(m-1)\sum_{i=1}^kn_i}$$
$$=1-(\frac{7*6+9*(\frac{7+9}{2})+8*(\frac{7+9+8}{3})+8*(\frac{7+9+8+8}{4})}{19*(7+9+8+8+8+8+7+8+8+6)}+\dots+\frac{8*(\frac{7+9+8+8+8+7+8+8}{8})+6*(\frac{7+9+8+8+8+7+8+8+6}{9})}{19*(7+9+8+8+8+8+7+8+8+6)})$$
$$=1-0.3861=0.6139$$

由计算结果得出 $\alpha_B(M_1)=0.6139>\alpha_B(M_0)=0.5933$ ，说明模型的行为适当性在添加低频事件日志序列 $\langle AB-CEPT\rangle$ 后得到了提高，从而加强了目标模型的行为适当性。

基于所得到的初始模型 M_0 和目标模型 M_1 ，根据算法 2 提出的模型加权值计算公式，得到表 4 的模型加权值表。为了更直观地观察结果，将得到的加权值用柱形图表示出来，具体见图 6(其中 W_i, W_k 分别为适合度和行为适当性的权重， Q_{M_0} 和 Q_{M_1} 表示在不同权重下模型 M_0 和 M_1 加权后的适合度以及行为适当性的值)。

表 4 模型加权值
Table 4 Weighted value of model

序列	W_k	W_i	Q_{M_0}	Q_{M_1}
1	0.40	0.60	0.9544	0.9595
2	0.45	0.55	0.9636	0.9689
3	0.50	0.50	0.9753	0.9783
4	0.55	0.45	0.9922	0.9987
5	0.60	0.40	0.9718	0.9871

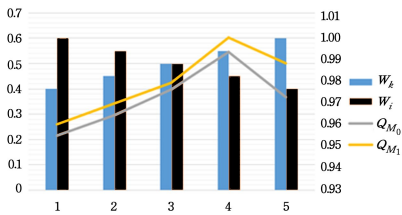


图 6 加权条形图

Fig. 6 Weighted bar chart

从表 4 中发现 $Q_{M_1}>Q_{M_0}$ ，且当 $W_k=0.55, W_i=0.45$ 时， Q_{M_1} 取最大值。利用 SPSS 软件对所得到的加权值进行曲线拟合处理，以进一步验证实验数据的准确性，具体结果如表 5、表 6 所列，拟合曲线如图 7 所示。对比表 5、表 6 发现， Q_{M_1} 的参数估计值大于 Q_{M_0} 的参数估计值，从图 7 中可以看出 Q_{M_1} 的拟合效果优于 Q_{M_0} ，因此模型 M_1 是我们最终挖掘得到的最优模型。

表 5 W_k 模型汇总和参数估计值

Table 5 Summary of W_k model and parameter estimation

方程	模型汇总					参数估计值		
	R 方	F	df_1	df_2	Sig.	常数	b_1	b_2
线性	0.504	3.046	1	3	0.179	-3.360	3.973	
二次	0.504	3.046	1	3	0.179	-3.360	3.973	0.000
复合	0.544	3.573	1	3	0.155	0.000	4293.145	
增长	0.544	3.573	1	3	0.155	-8.829	8.365	

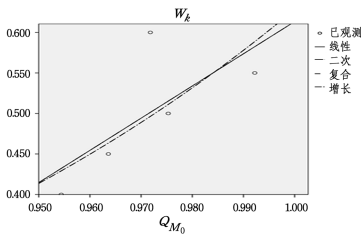
注：因变量为 W_k ，自变量为 Q_{M_0} 。

表 6 W_i 模型汇总和参数估计值

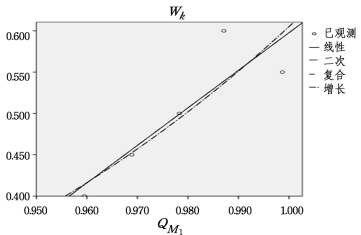
Table 6 Summary of W_i model and parameter estimation

方程	模型汇总					参数估计值		
	R 方	F	df_1	df_2	Sig.	常数	b_1	b_2
线性	0.773	10.190	1	3	0.050	-3.947	4.544	
二次	0.773	10.190	1	3	0.050	-3.947	4.544	0.000
复合	0.796	11.715	1	3	0.042	5.261×10^{-5}	11502.506	
增长	0.796	11.715	1	3	0.042	-9.853	9.350	

注：因变量为 W_i ，自变量为 Q_{M_1} 。



(a) Q_{M_0} 的曲线拟合



(b) Q_{M_1} 的曲线拟合

图 7 曲线拟合

Fig. 7 Curve fitting

结束语 本文以行为轮廓为基础,从行为角度考虑业务流程变化的部分,提出基于流程切的挖掘流程模型的方法。采用流程切的方式对事件日志进行分类,并基于行为轮廓构建初始模型。通过低频日志对初始模型进行补充和优化,通过行为适合度以及行为适当性等指标,对所建立的模型进行计算验证,以说明该方法的有效性。在日常生活中,业务流程所记录的事件日志数量都非常庞大,流程模型结构也复杂多样。在未来的工作中,基于复杂的业务流程系统,如何对过程模型进行挖掘,仍然是个难点;如何改进算法的性能以及改善算法的复杂性,将是研究的重点。

参 考 文 献

[1] VAN DER AALST W, WEIJTERS T, MARUSTER L. Workflow mining: discovering process models from event logs[J]. IEEE Transactions on Knowledge & Data Engineering, 2004, 16(9):1128-1142.

[2] VAN DER AALST W M P. Process Mining: Discovery, Conformance and Enhancement of Business Processes[M]. Springer Publishing Company, Incorporated, 2011.

[3] VAN DER AALST W M P, VAN DONGEN B F, HERBST J, et al. Workflow mining: a survey of issues and approaches[J]. Data & Knowledge Engineering, 2003, 47(2):237-267.

[4] VAN DER AALST W M P, WEIJTERS A. Process mining: a research agenda[J]. Computers in Industry, 2004, 53(3):231-244.

[5] CONFORTI R, ROSA M L, HOFSTEDE A H M T. Filtering Out Infrequent Behavior from Business Process Event Logs[J]. IEEE Transactions on Knowledge & Data Engineering, 2017, 29(2):300-314.

[6] EICHSTÄDT S, LINK A, HARRIS P, et al. Efficient implementation of a Monte Carlo method for uncertainty evaluation in dynamic measurements[J]. Metrologia, 2012, 49(3):401.

[7] POURMASOUMI A, KAHANI M, BAGHERI E. Mining variable fragments from process event logs[J]. Information Systems Frontiers, 2017, 19(6):1423-1443.

[8] WEN L, VAN DER AALST W M P, WANG J, et al. Mining process models with non-free-choice constructs[J]. Data Mining and Knowledge Discovery, 2007, 15(2):145-180.

[9] BUSI N, PINNA G, VAN DER AALST W. An Iterative Algorithm for Applying the Theory of Regions in Process Mining[R]. Department of Technology Management, Eindhoven University of Technology, 2007:16-38.

[10] YZQUIERDO-HERRERA R, SILVERIO-CASTRO R, LAZO-CORTÉS M. Sub-process discovery: opportunities for process diagnostics[M]//Enterprise Information Systems of the Future. Berlin Heidelberg: Springer, 2013:48-57.

[11] BOUSHABA S, KABBAJ M I, BAKKOURY Z. Process discovery: Automated approach for block discovery[C]//2014 International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE). IEEE, 2014:1-8.

[12] CHAPELA-CAMPA D, MUCIENTES M, LAMA M. Discovering Infrequent Behavioral Patterns in Process Models[C]//International Conference on Business Process Management. Springer, Cham, 2017:324-340.

[13] 吴哲辉. Petri 网理论[M]. 北京:机械工业出版社, 2006:6-22.

[14] WEIDLICH M, POLYVYANYYY A, DESAI N, et al. Process Compliance Measurement Based On Behavioral Profiles [J]. Advanced Information Systems Engineering, 2010, 6051:499-514.

[15] KUNZE M, WEIDLICH M, WESKEM. Querying process models by behavior inclusion[J]. Software & Systems Modeling, 2015, 14(3):1105-1125.

[16] WEIDLICH M, MENDLING J, WESKE M. Efficient Consistency Measurement Based on Behavioural Profiles of Process Models[J]. IEEE Transactions on Software Engineering, 2011, 37(3):410-429.