基于行为轮廓的业务流程隐变迁挖掘方法

宋 健 方贤文 王丽丽 刘祥伟

(安徽理工大学力学与光电物理学院 安徽 淮南 232001)

摘 要 在业务流程优化过程中,从非频繁行为中挖掘隐变迁是重要任务之一。从非频繁行为中挖掘隐变迁,能够更好地还原流程模型,提高流程的运行效率。文中依据行为轮廓的理论,在频率较高的日志中进行挖掘以获得初始模型。首先利用合理性阈值对事件日志进行过滤,得到有效的低频序列日志;其次利用低频序列日志优化初始模型,通过对各活动间行为轮廓关系与源模型的对比,来找到变化的区域,将可能存在的隐变迁挖掘出来;然后通过优化指标对挖掘到的隐变迁进行进一步验证,从而得到完整的含隐变迁的过程模型;最后通过具体的事例以及仿真对所构建的模型进行分析,并验证该方法的有效性。

关键词 行为轮廓,流程模型,隐变迁,Petri 网

中图法分类号 TP391.9 **文献标识码** A **DOI** 10.11896/jsjkx.180901654

Method of Mining Hidden Transition of Business Process Based on Behavior Profiles

SONG Jian FANG Xian-wen WANG Li-li LIU Xiang-wei

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

Abstract In the process of business process optimization, mining hidden transitions from infrequent behaviors is one of the important tasks. Mining the hidden transitions from infrequent behaviors can better restore the process model and improve the efficiency of the process. Based on the theory of behavioral profiles, this paper mined logs from relatively high frequency and obtained initial models. Firstly, the event log is filtered by the reasonableness threshold to obtain a valid low-frequency sequence log. Then, the low-frequency sequence log is used to optimize the initial model, and the behavioral profiles relationship between each activity is compared with the source model to find the changed region, and the possible hidden transitions are minned. Next, through the optimization of the indicators to further verify the hidden transitions, a complete and accurate model of the implicit transition process is obtained. Finally, the model is analyzed by concrete examples and simulations to verify the effectiveness of the method.

Keywords Behavioral profiles, Process model, Hide transition, Petri net

1 引言

随着信息技术的发展,业务流程管理被各行各业关注。流程挖掘的主要目的是从业务流程的事件日志中挖掘主要的信息,通过分析这些日志来检测该业务流程所存在的异常和偏差,以改善流程模型,从而达到优化的目的。然而,在实际的流程挖掘过程中,流程模型中总是存在一些这样的变迁,即在流程模型中是存在的,而在所执行的日志中并没有出现,这种变迁被称为隐变迁。为了更好地改善业务流程模型,从事件中挖掘隐变迁就变得尤为重要,通过挖掘到的隐变迁能够更好地还原模型,使业务流程模型更加完善、稳定。

目前,关于过程挖掘的方法众多,研究人员在过程挖掘方面也做了许多工作。文献[1]提出了一种启发式过程挖掘算法,它能够处理噪音,还能表示事件日志中的主要行为。文献[2]在过程模型的发现过程中,对非频繁行为(偏差或异常迹)进行搜索,并将其删除以降低模型的复杂性。文献[3]中的流程挖掘技术能够从现代信息系统普遍产生的事件中抽取信息,并建立清晰的流程模型,保证所构建的流程模型与实际的流程执行过程保持一致,以检测和改进实际过程。文献[4]提出了一种递归地探索候选过程树的方法,利用过程树来搜索本地流程模型。文献[5]提出了能够系统地处理生命周期信息的过程挖掘方法,以及一种能够处理生命周期数据并区分并发和交错的过程发现技术。文献[6]基于定位事件的挖掘算法,通过给每个事件分配一个非空的域集,来分析被定位的事件。文献[7]提出了局部过程模型(LPM)描述发生在较少结构化业务流程中的结构片段的过程行为,并且提出了一个

到稿日期:2018-09-05 返修日期:2019-01-10 本文受国家自然科学基金(61572035,61272153,61402011),安徽省自然科学基金(1508085MF111),安徽省高校自然科学基金(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;**刘祥伟**(1977-),女,副教授,主要研究方向为业务流程管理。

基于效用函数和约束条件的目标驱动的 LPM 发现框架。文献[8]提出了判定日志与模型一致性的分析方法,通过日志序列在模型中重放来计算其合理性和适当性。

本文通过事件日志活动间的行为轮廓关系来挖掘日志中的隐变迁。在流程挖掘之前,预先定义一个合理性阈值,对事件日志进行预处理操作。通过与合理性阈值的比较来挑选出事件日志中发生频率较高的日志,利用改组事件日志序列构建行为轮廓关系表,依据行为轮廓关系表建立源模型。对于小于合理性阈值的低频序列日志,构建低频序列日志的行为轮廓关系表,利用各活动间的行为轮廓关系优化源模型,找出模型结构中可能存在的隐变迁活动。利用精确度和查全率对疑似的隐变迁进行进一步的计算,过滤掉不合理的变迁活动,从而挖掘出含有隐变迁的目标优化模型。该方法所得到的过程模型更加准确、稳定,提高了模型的利用效率,实现了对流程模型的优化。

本文第 2 节介绍了基本概念;第 3 节提出基于行为轮廓挖掘隐变迁的方法;第 4 节给出了案例分析,并通过相关度量值进行计算,验证方法的可行性;第 5 节给出了实验仿真,对所得到的实验数据以及模型进行仿真,从而进行验证;最后总结全文并展望未来。

2　基本概念

定义 1[9](可行迹)　设一个流程模型 Petri 网 $BPP=(P,T;F,C)$,发生系列集合为 T_{pm},$\tau=n_1n_2\cdots n_k$,若$(x,y)\subseteq(N\cup F)\times(F\cup N)$,在 τ 中存在 $j\in(1,2,\cdots,k-1)$,$j<h\leqslant k$,有 $n_j=x$,$n_h=y$,则 τ 为一条可行迹,且有 $\tau\in T_{pm}$,记 $x<y$。

定义 2[10](弱序关系)　设(N,M_0)是一个网,M_0 是初始标识,若 $j\in(1,2,\cdots,n-1)$,$j<k\leqslant n$ 且存在一个发生序列 $\sigma=t_1,t_2,\cdots,t_n$ 使得$(N,M_0)[\sigma\rangle$,则对于所有活动变迁对$(x,y)\in T\times T$ 满足弱序关系,即 $t_1\succ t_j$。

定义 3[10](行为轮廓)　$PN=(P,T,F)$是一个 Petri 网,对任意的变迁对$(x,y)\in(T\times T)$,满足下列关系之一:

(1)严格序关系→,当且仅当 $x\succ y$ 且 $x\not\succ y$;

(2)排他性关系+,当且仅当 $x\not\succ y$ 且 $y\not\succ x$;

(3)交叉序关系$\|$,当且仅当 $x\succ y$ 且 $y\succ x$。

以上几种行为关系构成 Petri 网的行为轮廓。其中,←表示逆严格序关系。

定义 4[11](日志的弱行为轮廓)　设 L 是一条日志,对任意两个活动 $X,Y\in L$,X 与 Y 之间的关系是以下 3 种之一:

(1)严格序关系$\rightarrow_L$,若满足 $N(X,Y)\neq0\wedge N(Y,X)=0$,记作 $X\rightarrow_L Y$;

(2)交叉序关系$\|_L$,若满足 $N(X,Y)\neq0\wedge N(Y,X)\neq0$,记作 $X\|_L Y$;

(3)排他序关系$+_L$,若满足 $N(X,Y)=0\wedge N(Y,X)=0$,记作 $X+_L Y$。

我们称集合 $B_L=\{\rightarrow_L,\|_L,+_L\}$ 为日志 L 的行为轮廓。

注:对任意活动 X,Y,若有 $N(X,Y)=n$,则表示 X,Y 在日志中出现了 n 次。

定义 5[12](事件日志)　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 6[13](行为精确度和查全率)　设 σ 是一个事件日志的迹,$L(\sigma)$为迹 σ 在一个事件日志中所发生的次数,N_r 和 N_m 分别表示 Petri 网的参考模型和挖掘模型,C_r 和 C_m 分别表示 N_r 和 N_m 的因果关系,行为精确度和查全率的计算式分别为:

$$B_P(L,C_r,C_m)=\left(\sum_{\sigma\in L}\left(\frac{L(\sigma)}{|\sigma|}\times\sum_{i=0}^{|\sigma|-1}\frac{|Enabled(Cr,\sigma,i)\cap Enabled(Cm,\sigma,i)|}{|Enabled(C_m,\sigma,i)|}\right)\right)/\sum_{\sigma\in L}L(\sigma)$$

$$B_R(L,C_r,C_m)=\left(\sum_{\sigma\in L}\left(\frac{L(\sigma)}{|\sigma|}\times\sum_{i=0}^{|\sigma|-1}\frac{|Enabled(C_r,\sigma,i)\cap Enabled(C_m,\sigma,i)|}{|Enabled(C_r,\sigma,i)|}\right)\right)/\sum_{\sigma\in L}L(\sigma)$$

定义 7[13](结构精确度和查全率)　设 $Nr=(P_r,T_r,F_r)$为参考模型,$N_m=(P_m,T_m,F_m)$为挖掘到的模型,C_r 和 C_m 分别表示 N_r 和 N_m 的因果关系,结构精确度和查全率的计算公式分别为:

$$S_P(N_r,N_m)=\frac{|C_r\cap C_m|}{|C_m|}$$

$$S_R(N_r,N_m)=\frac{|C_r\cap C_m|}{|C_r|}$$

定义 8[14](低频模式)　设 L 是过程日志的迹集,一个简单模式 S_P 的频率为$freq(S_P)=\frac{|\{\tau\in L:S_P\ \alpha\ \tau\}|}{|L|}$,而模式 P 的频率为简单模式的最大频率,即 $freq(P)=\max\ freq(S_P)\ \forall S_P\in p$。给定一个频率阈值 $\sigma\in R:0<\sigma\leqslant1$,模式 P 是一个低频模式当且仅当 $freq(P)<\sigma$。

定义 9(隐变迁)　设 T'是 Petri 网业务流程模型中的变迁集,L'是记录日志事件集。$\lambda:T'\rightarrow L'$是标记映射,变迁 t'被称作隐变迁当且仅当 $t'\notin dom(\lambda)$,即变迁 t'不在 λ 的定义域内。

3　基于行为轮廓挖掘隐变迁的方法

隐变迁存在于流程模型中,但在事件日志中不存在,即日志中并没有显现出来。在实际的业务流程模型中,这样的活动变迁普遍存在,为了使流程模型更加完善,更加符合人们的生产需求,从事件日志中挖掘出隐变迁就显得尤为重要。

本文通过行为轮廓的理论来挖掘事件日志中的隐变迁。首先从大量流程日志中将高频的日志序列全部挖掘出来,基于这些高频日志以及各个变迁之间的行为轮廓关系,构建出业务流程的初始模型 M_0。利用行为轮廓关系,从低频序列中找出非噪音的日志序列进行补充,将其构建为模型 M_1。根据文献[15]所提出的适合度算法来对 M_1 模型进行计算,并且与原模型的适合度进行比较,若该片段对模型的适合度降低,则视为冗余,将其当作噪音过滤掉;若该片段对模型的适合度提高,则说明该低频是有效的。然后结合该低频序列的行为轮廓关系,查找出可能含有隐变迁的位置,并进一步依据行为的适当性进行验证。本文将所有的低频序列都按照此方法重复操作,保留对模型有改善的低频,过滤掉无用的低频,最终挖掘出含有隐变迁的目标模型。

根据给定的日志序列,需要对模型的结构适合度进行判断[15],适合度的具体定义如下:设 $L_P=n_1,n_2,\cdots,n_m$ 是流程模型 $P=(A,C,T,F)$的一组日志。集合 $SR\subseteq(AL\times AL)$包

含所有的活动变迁对(x,y)，其中日志 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 $\varepsilon_{LP}=\frac{|SR|}{|(AL\times AL)|}$（通常 $\varepsilon_{LP}\geqslant 0.8$，说明满足重放适合度）。再考虑模型的行为的适当性 α_B（所观察到的行为在此模型中的精确程度）：

$$a_B=1-\frac{\sum_{i=1}^{k}n_i(x_i-1)}{(m-1)\sum_{i=1}^{k}n_i}$$

其中，α_B（通常 $\alpha_B\geqslant 0.8$，说明满足行为的适当性）值越大，精确度将越高，挖掘的模型就越准确。假设 k 是聚合日志中不同轨迹的数量，n 为日志文件中所含的数目，m 为标记的数量任务（即不包括不可见任务，并且假设 $m>1$），x 为日志重放期间转换的平均数量。

算法 1 从事件日志中找出符合流程模型的低频序列

BEGIN(算法开始)

输入：事件日志序列 L，定义合理性阈值 t_f 和频率阈值 t_r(通常取0.1)

输出：符合流程模型的低频序列日志

步骤 1 将所得到的日志序列按照频率大小依次排列，eg.$\{t_1,t_2,t_3,\cdots,t_n\}$，$n\in\{1,2,3,\cdots\}$，$t_1,t_2,t_3,\cdots,t_n\in L$。

步骤 2 对日志进行预处理，将不完备的事件日志过滤掉，并将相同的序列日志进行合并。

步骤 3 记日志的总频数为 N，每条日志的频数分别记为 n_i(i=1,2,…,n)，计算日志的频率 $\sigma_i=\frac{n_i}{N}$，i=1,2,3…，若 $\sigma_i\geqslant t_r$，则归为频繁序列集，若 $\sigma_i<t_r$，则归为非频繁序列集。

步骤 4 根据定义 3 提出的行为轮廓定义，计算出各频繁序列集合中各个变迁之间的行为轮廓关系，并制作行为轮廓关系表。根据行为轮廓关系表构建初始模型 M_0。

步骤 5 在步骤 4 中构建的初始模型 M_0 并未考虑日志中的非频繁序列，因此得到的模型并不完善。为了使模型更加完整精确，需要将这些低频序列(非频繁序列集)考虑到初始模型 M_0 中。根据低频模式的定义，计算日志的频率 freq(i)，i=1,2,3,…。若 $freq(i)>t_f$，则该日志序列属于噪音序列日志，可以直接过滤掉；若 $freq(i)\leqslant t_f$，则该条日志是满足合理性的低频序列日志，将该序列保留，执行步骤(6)。

步骤 6 利用行为轮廓关系，将步骤 5 得到的日志序列重放到初始模型 M_0 中，对模型进行重新构建，假设经过添加新的序列所得到的模型为 M_1，利用第 3 部分提出的重放适合度定义，计算模型 M_0 和 M_1 的重放适合度值，分别记作 $\varepsilon_{LP(M0)}$ 和 $\varepsilon_{LP(M1)}$，若 $\varepsilon_{LP(M1)}\geqslant\varepsilon_{LP(M0)}$，说明将该序列日志重放至模型中使模型的适合度得到提高，则保留该日志序列，否则直接删除。

步骤 7 重复步骤 6，对所有的非频繁序列进行重放适合度计算，将所有 $\varepsilon_{LP(M)}\geqslant\varepsilon_{LP(M_0)}$(i=1,2,…)的日志都保留，否则直接过滤。

步骤 8 删除所有的低频序列模式的冗余日志信息，输出流程模型的低频序列日志。

END(算法结束)

算法 1 中，执行事件日志 L，将日志序列中的高频序列抽离出来。有些低频序列虽然属于低频，但在模型中是有意义的，并不属于噪音序列。因此，本文通过计算序列的合理度，来判断该序列的合理度是否满足我们所定义的要求。若满足预定义的阈值，则保留，否则直接删除。当所有的低频序列找到后，需要借助低频序列构建完善的模型，再从行为的适当性进行考虑，若满足行为适当性所设的阈值则保留，否则重新安放，若该序列对模型的行为度还是没有改善，则视为冗余删除。算法 2 给出了基于行为轮廓的隐变迁挖掘方法。

算法 2 基于行为轮廓的隐变迁挖掘方法

BEGIN(算法开始)

输入：算法 1 得到的事件日志序列，精确度 δ(通常介于 0.8 到 1.0 之间)

输出：挖掘出含有隐变迁的 Petri 网模型

步骤 1 将算法 1 所得到的日志序列按照频率大小依次排列，eg.$\{t_1,t_2,t_3,\cdots,t_n\}$，$n\in\{1,2,3,\cdots\}$，$t_1,t_2,t_3\cdots t_n\in L$。

步骤 2 算法 1 基于高频序列日志已经建立了初始模式 M_0。利用行为轮廓定义，建立出低频序列日志的行为轮廓关系，找出其变化区域。构建增量模块($M_1,M_2\cdots$)，将得到的增量模块配置到初始模型中，并执行步骤 3。

步骤 3 对于步骤 2 中找出的增量模块，利用行为轮廓关系找出变化区域部分可能存在的变迁对，再对照原事件日志序列的变迁对之间的关系。若存在变化，则说明该处存在疑似变迁，可能是隐变迁，也可能是其他因素(业务流程发生改变引起的变迁对变化)产生的，为了进一步验证，执行步骤 4。

步骤 4 将增量模块的序列日志重放至初始模型中，因为算法 1 已经对重放适合度进行了计算，因此当前需要计算模型的行为适当性，根据行为适当性计算公式 $\partial_B=1-\frac{\sum_{i=1}^{k}n_i(x_i-1)}{(m-1)\sum_{i=1}^{k}n_i}$，令重放后的模型为 M_i(i=1,2…)，若计算得出 $\partial_B(M_0)\geqslant\partial_B(M_1)$，则保留该增量日志模块，否则过滤处理。

步骤 5 考虑模型的适合度以及模型的行为适合度，分别设置权重 W_i 和 W_k(其中 $W_i\geqslant W_k$)。同时设置不同参数的权重 W_i，W_k，将其代入公式 $Q_M=\frac{W_i\varepsilon_{LP}+W_k\partial_B}{W_i+W_k}$ 中，若 $Q_{M1}>Q_{M0}$，则得出最优值，执行步骤(6)。

步骤 6 根据步骤 4 得到的序列日志(包含疑似隐变迁的序列)继续构建新的增量子模块，将可能含有隐变迁的活动标记出来，将挖掘到的隐变迁放置到初始模型中，对初始模型进行补充，得到目标模型 M_1。根据定义[6]，挖掘得到的模型 M_1 需要考虑模型的行为精确度 $B_P(L,C_{M0},C_{M1})$ 和行为查全率 $B_R(L,C_{M0},C_{M1})$，若 $B_P\geqslant\delta\&\&B_R\geqslant\delta$，则所挖掘到的模型在行为上是符合语义的，否则不符合行为语义，需要进行过滤。

步骤 7 步骤 5 完成后，依据定义[7]，需要再次计算模型的结构精确度 $S_P(N_{M0},N_{M1})$ 和结构查全率 $S_R(N_{M0},N_{M1})$，若 $S_R\geqslant\delta\&\&S_P\geqslant\delta$，则所挖掘到的模型在结构上符合要求，否则不符合结构要求，将其过滤掉。

步骤 8 经步骤 7，所得到的变迁为最终满足要求的变迁——隐变迁，模型为最终包含隐变迁的目标模型。最后输出符合要求的包含隐变迁 Petri 网的优化模型。

END(算法结束)

4 案例分析

为了验证算法的准确性，给出了一个超市购物的实例，以验证算法的可行性。为方便区分各个活动，采用不同的字母来代表各个活动，其中顾客活动用大写字母进行表示，商家活动用小写字母进行表示。事件日志中各字母所代表的具体信

息如下:A 表示进店选物品、B 表示排队付款、C 表示选择现金支付、D 表示选择购物卡、E 表示网银支付、F 表示确认支付、G 表示放弃支付、H 表示余额不足、I 表示余额充足、J 表示确认支付、K 表示确认支付、L 表示收款成功、M 表示打印发票、N 表示交易结束、a 表示对商品扫码、b 表示统计价格、c 表示收银机统计价格、d 表示待付款。具体的事件日志如表 1 所列。

表 1　执行日志序列 L

Table 1　Execution log sequence L

日志名称	事件日志轨迹	实例数
L1	*AabcdBCJLMN*	983
L2	*ABabcdCJLMN*	1623
L3	*ABabcdDIKLMN*	1238
L4	*ABabcdEFLMN*	1129
L5	*ABabcdDHGCJLMN*	980
L6	*ABCJLMN*	23
L7	*ABabcdDHGEFLMN*	898
L8	*abcdLMN*	18
L9	*abcd*	10
L10	*AabcdBDHGCHLMN*	83
L11	*AabcdBDHGEHLMN*	98
L12	*AabcdBDIKLMN*	883
L13	*AabcdBEFLMN*	766
L14	*AabcdBDHGCJLMN*	839
L15	*AabcdBDHGEFLMN*	689
L16	*ABabcd*	9
L17	*ABab*	8
L18	*ABabcdDHGCHLMN*	78
L19	*ABabcdDHGEHLMN*	117
L20	*AB*	2
L21	…	…

将表 1 中的日志序列由实例数从高到低顺序为:{〈*ABabcdCJLMN*〉^{1623}、〈*ABabcdDIKLMN*〉^{1238}、〈*ABabcdEFLMN*〉^{1129}、〈*AabcdBCJLMN*〉^{983}、〈*ABabcdDHGCJLMN*〉^{980}、〈*ABabcdDHGEFLMN*〉^{898}、〈*AabcdBDIKLMN*〉^{883}、〈*AabcdBDHGCJLMN*〉^{839}、〈*AabcdBEFLMN*〉^{766}、〈*AabcdBDHGEFLMN*〉^{689}}。根据定义 5,建立活动关系表,如表 2 所列,再根据定义 4 制作日志序列的行为轮廓关系表,具体如表 3 所列。

通过表 2 和表 3 的活动关系以及行为轮廓关系表之间的关系,来建立初始模型 M_0,如图 1 所示。

表 2　日志序列活动关系表

Table 2　Log sequence activity relationship table

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	a	b	c	d
A	0	5	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0
B	0	0	1	3	1	0	0	0	0	0	0	0	0	0	5	0	0	0
C	0	0	0	0	0	0	0	4	2	4	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	4	2	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0
G	0	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	1	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0
b	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0
c	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10
d	0	5	1	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0

表 3　日志序列行为轮廓关系表

Table 3　Log sequence behavior profile table

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	a	b	c	d
A	+	→	→	→	→	→	→	→	→	→	→	→	→	→	‖	‖	‖	‖
B	←	+	→	→	→	→	→	→	→	→	→	→	→	→	‖	‖	‖	‖
C	←	←	+	+	+	+	+	+	+	→	+	→	→	→	‖	‖	‖	‖
D	←	←	+	+	+	+	→	→	→	+	→	→	→	→	‖	‖	‖	‖
E	←	←	+	+	+	→	+	+	+	+	→	→	→	→	‖	‖	‖	‖
F	←	←	+	+	←	+	+	+	+	+	→	→	→	→	‖	‖	‖	‖
G	←	←	+	←	←	←	+	←	+	+	→	→	→	→	‖	‖	‖	‖
H	←	←	+	+	+	+	+	+	+	+	→	→	→	→	‖	‖	‖	‖
I	←	←	+	+	+	+	+	+	+	→	→	→	→	→	‖	‖	‖	‖
J	←	←	+	+	+	+	+	+	←	+	→	→	→	→	‖	‖	‖	‖
K	←	←	+	+	+	+	+	+	←	←	+	→	→	→	‖	‖	‖	‖
L	←	←	+	+	+	+	+	+	←	←	←	+	→	→	‖	‖	‖	‖
M	←	←	+	+	+	+	+	+	←	←	←	←	+	→	‖	‖	‖	‖
N	←	←	+	+	+	+	+	+	←	←	←	←	←	+	‖	‖	‖	‖
a	‖	‖	‖	‖	‖	‖	‖	‖	‖	‖	‖	‖	‖	‖	+	→	→	→
b	‖	‖	‖	‖	‖	‖	‖	‖	‖	‖	‖	‖	‖	‖	←	+	→	→
c	‖	‖	‖	‖	‖	‖	‖	‖	‖	‖	‖	‖	‖	‖	←	←	+	→
d	‖	‖	‖	‖	‖	‖	‖	‖	‖	‖	‖	‖	‖	‖	←	←	←	+

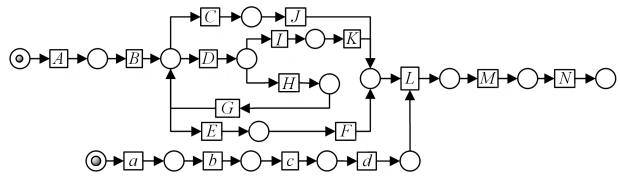


图 1　Petri 网的业务流程初始模型 M_0

Fig. 1　Petri net business process initial model M_0

根据文献[14]提出的合理性概念,本文得到的初始模型 M_0 是合理的,对于其中不符合实际运行的日志,如日志〈*AB*〉,〈*abcd*〉,〈*ABab*〉等,因为这些日志是其他因素产生的,对构建模型没有任何帮助,而且会降低模型的适合度和行为适当性,所以可以直接将其视为噪音过滤掉。通过适合度公式的计算得到 $\varepsilon_{lp0}>0.901$,再通过行为适当性计算得到 $\partial_{B0}>0.901$。日志 L_{10},L_{11},L_{18},L_{19} 并未参与初始模型 M_0 的构建,因此在计算模型的合适度以及行为适当性过程中该组日志(L_{10},L_{11},L_{18},L_{19})并未被考虑。在接下来挖掘隐变迁的过程中,依据算法 2,利用这些低频日志对模型进行补充完善。分析序列日志 L_{10},L_{18} 可以发现,在该日志付款流程中都存在 $D\rightarrow H\rightarrow G\rightarrow C\rightarrow H\rightarrow L\rightarrow M\rightarrow N$ 这样的序列。在所构建的初始模型中,通过分析这些日志可以发现,虽然日志的频率相对较低,但其对模型的稳定性以及适合度方面都有改善。本文对日志 L_{10},L_{18} 进行分析,挖掘带有隐变迁的子模块图 M_1,得到的隐变迁用字母 O 表示。挖掘到子模块后,将其放入初始流程图中,并通过行为适当性计算 $\partial_{B1}=1-\frac{14*13+14*13}{(83+78)*(14+14)}=0.9193>\partial_{B0}=0.901$,因此该子模块是有效的,日志序列 L_{10},L_{18} 被视为有效低频,子模块 M_1 如图 2 所示。同理,对日志 L_{11},L_{19} 进行同样的分析,并且通过行为适当性计算 $\partial_{B2}=1-\frac{14*13+14*13}{(98+117)*(14+14)}=0.9395>\partial_{B1}=0.9193>\partial_{B0}=0.901$,发现行为适当性得到进一步提高,因此日志序列 L_{11},L_{19} 也是有效的低频序列日志,挖掘出带有隐变迁的子模块 M_2,得到的隐变迁用 P 表示,子模块 M_2 如图 3 所示。

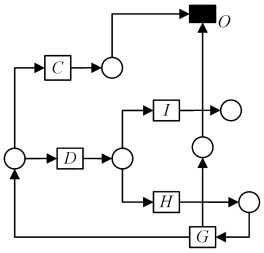


图 2 带有隐变迁的子模块 M_1

Fig. 2 Module with hidden transition M_1

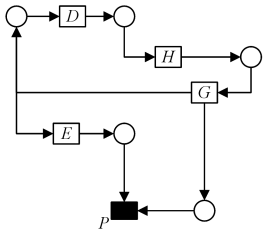


图 3 带有隐变迁的子模块 M_2

Fig. 3 Module with hidden transition M_2

最后将子模块 M_1 和 M_2 合并到初始模型 M_0 中，对初始模型进行补充，并完善初始模型，最终得到图 5 所示的含有隐变迁的模型 M_3。本文设置参数 W_i，W_k，其分别为适合度和行为适当性的权重，根据算法 2 计算不同权重下 Q_{M_0}，Q_{M_3} 的值，具体计算结果如表 4 所列。

表 4 模型加权值表

Table 4 Model weighting table

序列	W_k	W_i	Q_{M_0}	Q_{M_3}
1	0.35	0.65	0.9059	0.9142
2	0.40	0.60	0.9066	0.9154
3	0.45	0.55	0.9073	0.9165
4	0.50	0.50	0.9081	0.9177
5	0.55	0.45	0.9087	0.9189
6	0.60	0.40	0.9084	0.9174

通过表 4 发现 $Q_{M_0}<Q_{M_3}$，即当 $W_k=0.45$，$W_i=0.55$ 时，Q_{M_3} 的值最大，因此 M_3 的模型即为我们所要的模型。在完善后的模型中，我们分析所挖掘到的隐变迁 O 和 P，可以知道该活动所表示的意义是当顾客利用购物卡进行付款时，顾客并没有直接选择放弃支付。顾客先用购物卡已有的现金进行付款，余下的部分将采用现金付款或网银支付，这样促使支付变得更加便捷，同时也使顾客的利益得到保障。完整的含有隐变迁的流程模型图如图 4 所示。

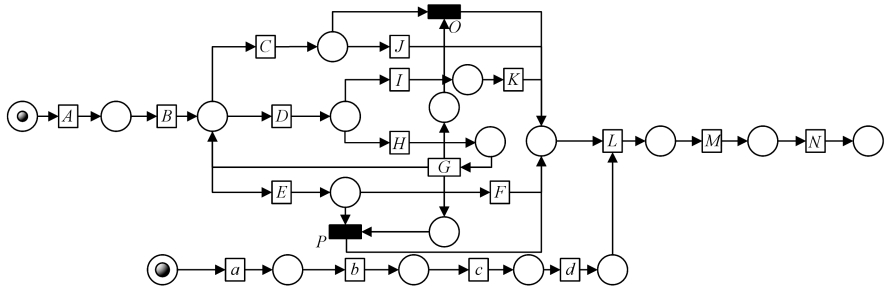


图 4 带有隐变迁的 Petri 网业务流程模型 M_3

Fig. 4 Petri net business process model with hidden transition M_3

根据定义 6 提出的概念以及算法 2 中的步骤 6，需要计算初始模型 M_0 和目标模型 M_3 的行为精确度和行为查全率（本文中的精确度 δ 取值为 0.85）：

$$B_P(L,C_{M_0},C_{M_3})=\frac{1}{376}*(\frac{83}{14}*(\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{3}{3}+\frac{1}{1}+\frac{1}{1}+\frac{2}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1})+\frac{98}{14}*(\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{3}{3}+\frac{1}{1}+\frac{1}{1}+\frac{2}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1})+\frac{78}{14}*(\frac{1}{1}+\frac{3}{3}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{2}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1})+\frac{117}{14}*(\frac{1}{1}+\frac{3}{3}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{2}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}))$$

$$=0.9643$$

同理算出：

$$B_R(L,C_{M_0},C_{M_3})=\frac{1}{376}*(\frac{83}{14}*(\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{3}{3}+\frac{1}{1}+\frac{1}{1}+\frac{2}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1})+\frac{98}{14}*(\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{3}{3}+\frac{1}{1}+\frac{1}{1}+\frac{2}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1})+\frac{78}{14}*(\frac{1}{1}+\frac{3}{3}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{2}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1})+\frac{117}{14}*(\frac{1}{1}+\frac{3}{3}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{2}{2}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}))$$

$$=1$$

通过计算得出 $B_P(L,C_{M_0},C_{M_3})=0.9643>0.85\&\&B_R(L,C_{M_0},C_{M_3})=1>0.85$，因此目标模型 M_3 在行为精确度和查全率上都满足要求。再通过算法 2 中的步骤 7 对初始模型 M_0 和目标模型 M_3 的结构精确度和结构查全率进行计算验证：

$$S_P(N_r,N_m)=|\{(A,B),(B,C)(B,D),(B,E),(C,J),(D,I),(D,H),(E,F),(F,L),(G,P),(G,C),(G,E),(H,G),(I,K),(J,L),(K,L),(L,M),(M,N),(a,b),(b,c),(c,d),(d,l)\}|/|\{(A,B),(B,C)(B,D),(B,E),(C,J),(C,O),(D,I),(D,H),(E,F),(E,P),(F,L),(G,O),(G,P),(G,C),(G,E),(H,G),(I,K),(J,L),(K,L),(L,M),(M,N),(a,b),(b,c),(c,d),(d,l)\}|=0.88$$

$$S_R(N_r,N_m)=|\{(A,B),(B,C)(B,D),(B,E),(C,J),(D,I),(D,H),(E,F),(F,L),(G,C),(G,E),(H,G),(I,K),(J,L),(K,L),(L,M),(M,N),(a,b),(b,c),(c,d),(d,l)\}|/|\{(A,B),(B,C)(B,D),(B,E),(C,J),(D,I),$$

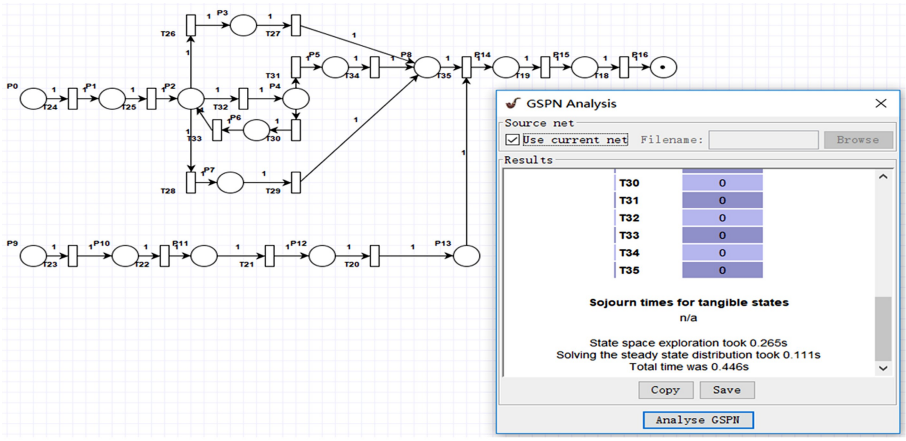
$(D,H),(E,F),(F,L),(G,C),(G,E),(H,G),(I,K),(J,L),(K,L),(L,M),(M,N),(a,b),(b,c),(c,d),(d,l)\}|=1$

通过计算得出 $S_P(N_r,N_m)=0.88>0.85\&\&S_R(N_r,N_m)=1>0.85$,可知目标模型 M_3 在结构精确度和结构查全率上都满足要求,因此含有隐变迁活动 O 和 P 的模型 M_3 即为所得到的最终模型。

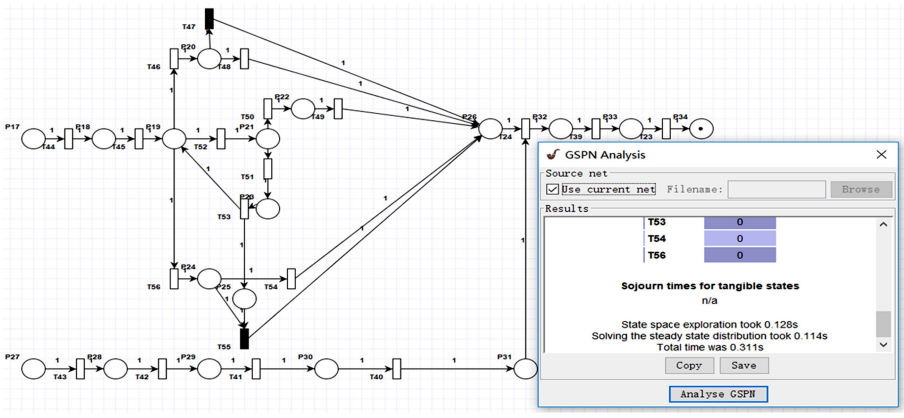
通过算法 1、算法 2 以及日志之间的行为轮廓关系,将隐变迁挖掘出来。隐变迁的挖掘使模型更加完整,模型稳定性也得到了提高。将隐变迁嵌入到初始模型中后,通过计算发现流程模型的行为精确度以及适当性也得到了很大的提高,使模型得到优化,更符合日志序列的要求。

5　仿真实验

为验证算法的有效性,采用 PIPE、SPSS 软件对挖掘的流程模型进行模型分析和数据分析。在 PIPE 软件中,将挖掘的源模型 Petri 网以及含隐变迁的目标模型 Petri 网分别进行仿真实验。将图 5(a)中标记为黑色的 token 点最终移动到 P16 库所中,说明整个 Petri 网是流通的、有界的,同理图 5(b)也同样得到类似的验证。对比图 5(a)和图 5(b)中的 GSPN 仿真结果可得,挖掘到含隐变迁的 Petri 网模型所消耗的总时间(0.311 s)优于源模型所消耗的总时间(0.446 s)。在表 5、表 6 中,W_i 和 W_k 通过 SPSS 软件对数据进行拟合处理,得到表 5、表 6 所列的参数估计值。从表 5、表 6 中发现,在参数估计值数据处理中,$W_i=0.913>W_k=0.902$,利用 SPSS 软件对生成的数据采用曲线拟合的方法进行处理,生成图 6 所示的曲线拟合,从图中能够更为直观地看出 $Q_{M_3}>Q_{M_0}$,而且图 6(b)所示的曲线拟合效果优于图 6(a),说明挖掘得到的业务流程 M_3 更符合我们的需求,模型也更加完善。



(a)源模型仿真图



(b)含隐变迁的目标模型仿真图

图 5　模型仿真图

Fig. 5　Model simulation diagram

表 5　自变量为 W_i、因变量为 Q_{M_0} 时的模型汇总和参数估计值

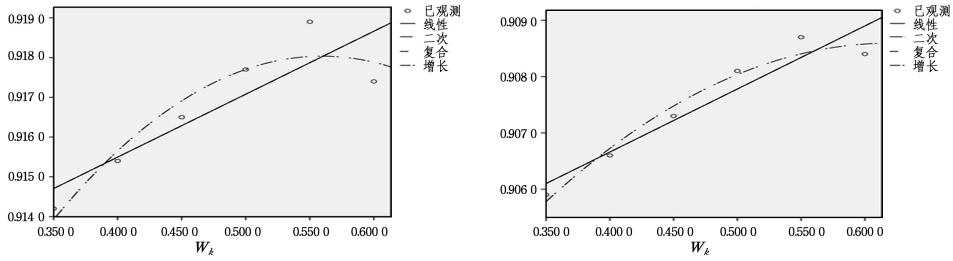
Table 5　Model summary and parameter estimates for independent variables W_i and dependent variables Q_{M_0}

方程	模型汇总					参数估计值		
	R 方	F	df1	df2	Sig.	常数	b1	b2
线性	0.766	13.100	1	4	0.022	−43.891	48.398	
二次	0.766	13.100	1	4	0.022	−43.891	48.398	0.000
复合	0.810	17.104	1	4	0.014	1.013×10^{-43}	3.479×10^{46}	
增长	0.810	17.104	1	4	0.014	−98.998	107.166	

表 6 自变量为 W_k、因变量为 Q_{M_3} 时的模型汇总和参数估计值

Table 6 Model summary and parameter estimates for independent variables W_k and dependent variables Q_{M_3}

方程	模型汇总					参数估计值		
	R 方	F	df1	df2	Sig.	常数	b1	b2
线性	0.912	41.263	1	4	0.003	−73.391	81.395	
二次	0.912	41.276	1	4	0.003	−36.466	0.000	44.855
复合	0.940	62.477	1	4	0.001	3.545×10^{-71}	1.850×10^{77}	
增长	0.940	62.477	1	4	0.001	−162.218	177.914	



(a)Q_{M_3} 的曲线拟合 (b)Q_{M_3} 的曲线拟合

图 6 模型曲线拟合图

Fig. 6 Model curve fitting diagram

结束语 本文在现有研究的基础上，提出基于行为轮廓从事件日志中挖掘隐变迁的方法。从事件日志中将符合业务流程的高频日志筛选出来，利用高频日志构建出初始模型。然后从低频事件日志中将不符合业务流程的事件日志删除，保留余下的低频日志，通过这些低频事件日志对模型进行进一步的优化和补充，挖掘出含有隐变迁的子模块，最后将挖掘到的子模块嵌入到初始模型中，对初始模型进行完善。通过计算行为的适当性和模式适合度，发现构建的模型在优化指标上有很大的提高，通过软件对所建的模型进行仿真分析，最后结合实例验证了该方法的有效性。

基于行为轮廓对隐变迁挖掘时，并没有将复杂模型中含有的隐变迁考虑在内，这是因为在复杂模型的系统中，挖掘隐变迁还是非常困难的。在未来的研究中，需要对复杂流程模型进行挖掘，将流程模型中的隐变迁挖掘出来，以满足人们对业务流程发展的需求，使流程挖掘技术得到更好的完善和发展。

参 考 文 献

[1] WEIJTERS A J M M, VAN DER AALST W M P, DE MEDEIROS A K A. Process mining with the heuristics miner-algorithm [R]. Technische Universiteit Eindhoven, Technical report WP 166, 2006: 1-34.

[2] CONFORTI R, ROSA M L, HOFSTEDE A H M. Filtering out infrequent behavior from business process event logs[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 29: 300-314.

[3] VAN DER AALST W M P. Process Mining Discovery, Conformance and Enhancement of Business Processes[M]. Berlin: Springer-Verlag, 2011: 191-211.

[4] BUIJS J C A M, VAN DONGEN B F, VAN DER AALST W M P. A genetic algorithm for discovering process trees[C]// Proceedings of the 2012 IEEE Congress on Evolutionary Computation (CEC), IEEE, 2012: 1-8.

[5] LEEMANS S J J, FAHLAND D, AALST W M P V D. Using Life Cycle Information in Process Discovery[M]// Business Process Management Workshops. 2016: 1-12.

[6] AALST W M P V D, KALENKOVA A, RUBIN V, et al. Process Discovery Using Localized Events [M]. Application and Theory of Petri Nets and Concurrency. Springer International Publishing, 2015: 287-308.

[7] NIEK T, BENJAMIN D, NATALIA S, et al. Interest-Driven Discovery of Local Process Models[J]. Information Systems, 2018: 1-16. arXiv: 1703.07116.

[8] ROZINAT A, W M PVAN DER AALST. Conformance Testing: Measuring the Fit and Appropriateness of Event Logs and Process Models[J]. Computer Science, 2005, 3812 LNCS: 163-176.

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

[10] FANG X W, WU J Z, LIU X W. An Optimized Method of Business Process Mining Based on the Behavior Profile of Petri Net [J]. Information Technology Journal, 2014, 13(1): 86-93.

[11] WEIDLICH M, WESKE M, MENDLING J. Change propagation in process models using behavioural profiles[C]// 2009 IEEE International Conference on Services Computing. Bangalore, India, 2009: 32-40.

[12] VAN DER AALST W, 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.

[13] CHENG H J, KUMARA. Process Mining on Noisy Logs—Can log sanitization help to improve performance? [J]. Decision Support Systems, 2015, 79(C): 138-149.

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

[15] 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.