基于 Petri 网行为紧密度的有效低频行为模式分析

郝惠晶 方贤文 王丽丽 刘祥伟

(安徽理工大学数学与大数据学院 安徽 淮南 232001)

摘 要 低频行为模式分析是流程管理的重要内容之一,有效区分低频日志和噪音日志在业务流程过程挖掘中显得尤为重要。目前已有的研究大部分是将流程模型中的低频行为当作噪音直接过滤,但有些低频行为对模型是有效的。文中提出了基于 Petri 网行为紧密度的有效低频模式分析方法。首先,根据给定的事件日志建立合理的流程模型;然后,通过迭代扩展初始模式来发现流程模型中的所有低频日志序列,并在此基础上计算日志与模型的行为距离向量,利用日志与模型的行为紧密度找出有效的低频行为模式;最后,通过实例分析验证了所提方法的可行性。

关键词 事件日志,低频模式,阈值,行为紧密度

中图法分类号 TP391.9 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.02.049

Analysis of Effective Low-frequency Behavioral Patterns Based on Petri Net Behavior Closeness

HAO Hui-jing FANG Xian-wen WANG Li-li LIU Xiang-wei

(College of Mathematics and Big Data, Anhui University of Science and Technology, Huainan, Anhui 232001, China)

Abstract Low-frequency behavior pattern analysis is one of the important contents of process management. It is very important to distinguish low-frequency logs and noise logs effectively in business process mining. At present, most of the researches have dealt with the direct filtering of low-frequency behavior in the process model as noise, but some low frequency behavior are valid for the model. This paper presented an effective low-frequency pattern analysis method based on Petri nets' behavioral closeness. Firstly, a reasonable process model is established according to the given event log. Then, all low-frequency log sequences in the process model are found by iteratively expanding the initial patterns. Based on this, the behavioral distance vectors of the log and the model are calculated, and the behavior closeness of log and model is used to find the effective low-frequency behavioral pattern. Finally, an example is given to verify the feasibility of this method.

Keywords Event log, Low frequency patterns, Threshold, Behavior closeness

1 引言

在大数据背景下,业务流程管理在很多领域都占据着不可忽视的地位。过程挖掘作为业务流程管理的重要内容之一,目的是从运行系统的事件日志中获取有用信息,从而建立合适的流程模型。近年来,过程挖掘的重点之一是研究发现频繁行为,并对常见行为进行发现、监测及应用,目的是发现运行示例中主要发生的过程,很少有工作涉及低频行为,并且已有工作主要集中在异常或偏差的检测上。但是低频行为对流程管理也很重要,它揭示了一个过程中低频率的错误实现,因此从过程模型中发现低频行为已成为当前学者研究的热门课题。

在过程挖掘领域中,从事件日志中推导流程模型被称为过程发现。已有研究主要是在过程中发现频繁行为,如文献[1-2]就提出了相关的算法来发现过程模型中的频繁行为。文献[3]提出了一种在事件日志中发现频繁行为模式的方法,这种方法涉及序列的组合、并发、选择和循环的行为模式,并且基于过程树捕获频繁模式来建立局部流程模型,对于一个给定的事件日志,对局部流程模型建立相应的质量维度和指标。但是事件日志通常也会包含一些低频行为或噪声,如不常见的异常值或错误记录,从而掩盖过程的主要行为。为了更好地研究业务流程,许多学者开始深入研究过程模型中的低频行为。文献[4]发现了真实事件日志中存在异常值或者噪音,这些噪音将导致数据输入错误和丢失;并且在过程发现中,低频事件也会导致不可用或复杂的模型。文献[5]在流程模型发现中搜索到一些低频行为,如偏差或异常迹,然后将其

到稿日期:2018-01-17 返修日期:2018-03-29 本文受国家自然科学基金项目(61572035,61402011),安徽省自然科学基金(1508085MF111,1608085QF149),安徽省高校自然科学基金重点项目(KJ2016A208),安徽理工大学研究生创新基金项目(2017CX2113)资助。

郝惠晶(1993-),女,硕士生,主要研究方向为 Petri 网,E-mail:609795483@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,硕士,副教授,主要研究方向为业务流程分析和软件认证;刘祥伟(1977-),女,硕士,教授,主要研究方向为业务流程管理。

删除以降低模型的复杂度。文献[6]提出了一种新的过程发现方法,即数据感知启发式挖掘,通过采用分类技术来区分不常见路径,主要从随机噪音中利用数据属性来区分,数据流和控制流在这个过程中会被同时发现。文献[7]提出了一种基于最大模式挖掘(MPM)进行过程发现的新方法,其根据整个事件序列中的可见迹来构建模式,以确保挖掘模型的合理性。该方法可以处理任何长度的循环、重复、无选择结构和距离依赖。同时,文献[7]基于频率提出了噪音过滤的方法。文献[8]通过在执行的流程示例中识别频繁行为和低频行为来检测偏差,并在 ProM 中对人造日志和真实日志进行评估。但是,并不是所有的低频行为在流程中都表现为噪音,有的低频行为可能会监测并优化过程,因此直接将其删除不是一个合适的解决方案。例如,文献[9]提出了一种能够处理低频行为和大量事件日志的新技术,在示例中返回的流程模型可以确保其合理性,没有死锁和其他异常出现,且能在 ProM 中进行实现。文献[10]提出了一种 WoMine-i 算法用于从过程模型中检索出低频行为模式,该算法可以在过程模型中搜索选择、并行、循环的序列,但这些结构在日志中很少被执行。

已有的发现低频行为的研究大多集中在对噪音及异常行为进行处理等方面,较少关注低频行为的可用性,这在实际应用中存在一定的局限性。针对这一问题,本文主要根据给定的事件日志找出合理的流程模型,并基于流程模型发现低频行为,然后根据行为紧密度的定义找出符合模型行为关系的有效低频行为模式。

本文第 2 节给出一个动机例子。第 3 节介绍行为轮廓及流程模型的相关知识。第 4 节重点阐述基于行为紧密度检测发现有效低频行为模式的方法,其中 4.1 节介绍基于事件日志建立合理的流程模型,然后发现低频行为模式;4.2 节提出用行为紧密度检测来发现有效的低频行为模式,并给出相关算法。第 5 节通过实例分析来验证算法的可行性。最后总结全文并展望未来的工作。

2 动机例子

当今,汽车已成为人们必不可少的交通工具,但也会导致各种交通事故,因此车辆的保险理赔就变得尤为重要。当前大部分的车辆保险理赔流程主要分为以下步骤:报案、查勘、定损、签收审核索赔单证、理算复核、审批、赔付结案等。但是由于保险理赔系统不够完善,可能会使保险公司蒙受不必要的损失,比如顾客欺诈性索赔、合同期间未按时缴费、出险车辆不满足理赔条件等。这些现象虽然在车辆保险理赔中不常见,但是至少说明保险公司的理赔系统还需要进一步完善。表 2 基于表 1 的说明列出了一些从理赔系统中调取的事件日志。

表 1 各个字母代表的活动事件

Table 1 Activity events represented by letters

活动	活动名称	活动	活动名称
A	拨打保险公司客服电话	J	拖车到保险公司定损点定损
B	填写出险登记表	K	现场勘查定损
C	查验保险	L	定损完成
D	确认保险合同是否有效	M	提交理赔材料
E	确认保险期限是否未满	N	审核并计算索赔
F	审核完成	O	确认赔付金额
G	确认出险	P	领取赔款
H	不用现场勘查	Q	交易关闭
I	保险公司专业人员定损	R	不满足理赔条件

表 2 车辆保险理赔的事件日志

Table 2 Event log of vehicle insurance claims

案例	事件日志	实例数
1	ABCDEFQ	15
2	ABCDEFGKLRQ	23
3	ABCDEFGHILRQ	13
4	ABCEDFGHJLRQ	18
5	ABCDEFGKLMNOPQ	2904
6	ABCEDFGHILMNOPQ	2288
7	ABCDEFGHJLMNOPQ	2556
8	ABCDEFGKLMNRQ	43
9	ABCEDFGHILMNRQ	28
10	ABCDEFGHJLMNRQ	35
11	ABCEDFGKLMNMNOPQ	734
12	ABCDEFGHILMNMNOPQ	648
13	ABCEDFGHJLMNMNOPQ	695

根据上述事件日志信息,构建了如图 1 所示的初始模型 M_0。该模型符合一般的车辆保险理赔程序。但是通过分析可以发现,实例数较少的日志在该模型中未被体现。本文即是针对实例数较少的日志进行研究,通过日志挖掘出合理的模型,找出模型中的低频行为模式,并分析日志与模型的关系,判断低频模式是有效低频还是噪音。

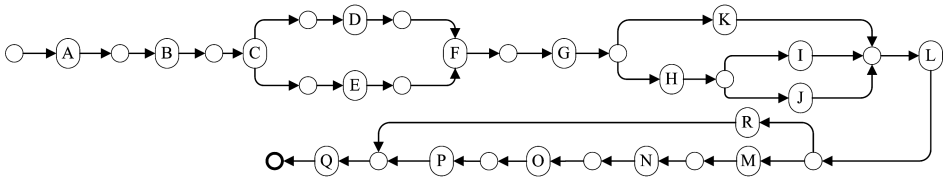


图 1 车辆保险理赔 Petri 网的初始模型 M_0

Fig. 1 Initial model M_0 of vehicle insurance claims Petri net

3 基础知识

定义 1[11](行为轮廓) 设(N,M_0)是一个网系统,其中$N=(P,T,F)$,初始标识为 M_0。对于任意给定的变迁对$(t_1,t_2)\in(T\times T)$:

1)若 $t_1\succ t_2$ 且 $t_2\not\succ t_1$,则称其为严格序关系,记作 $t_1\rightarrow t_2$;

2)若 $t_1\not\succ t_2$ 且 $t_2\succ t_1$,则称其为严格逆序关系,记作 $t_1\rightarrow^{-1} t_2$;

3)若 $t_1\not\succ t_2$ 且 $t_2\not\succ t_1$,则称其为排他关系,记作 t_1+t_2;

4)若 $t_1\succ t_2$ 且 $t_2\succ t_1$,则称其为交叉序关系,记作 $t_1\parallel t_2$。

将所有关系的集合叫做网系统的行为轮廓,则称 $BP=\{\rightarrow,\rightarrow^{-1},+,\parallel\}$是网系统的行为轮廓。

定义 2[12](业务流程模型的 Petri 网)　一个流程模型的 Petri 网是一个四元组 $N=(P,T,F,C)$。其满足以下几个条件:

1)P 为有限非空库所的集合,T 为有限非空变迁的集合,$P\cap T=\emptyset$。

2)$F\subseteq(P\times T)\cup(T\times P)$为 N 中的流关系。

3)$\cdot x=\{y\mid(y\in P\cup T)\cap((y,x)\in F)\}$为 x 的前集,$x\cdot=\{y\mid(y\in P\cup T)\cap((y,x)\in F)\}$为 x 的后集。

4)若 x 为 P 的初始库所,则 $\cdot x=\emptyset$;若 x 为 P 的终止库所,则 $x\cdot=\emptyset$;$dom(F)\cup cod(F)=P\cup T$ 且 $dom(F)=\{x\in P\cup T,\exists y\in P\cup T,(x,y)\in F\}$,$cod(F)=\{x\in P\cup T,\exists y\in P\cup T,(y,x)\in F\}$。

5)$C=\{od,or,pl,cy\}$为业务流程 Petri 网的结构类型,即包括顺序、选择、并行和循环 4 种结构。

定义 3[6](因果网)　一个因果网(C-网)是一个元组 $C=(A,a_i,a_o,D,I,O)$,且 A 是一个有限的活动集;$a_i\in A$ 是开始活动,$a_o\in A$ 是结束活动;$D\subseteq A\times A$ 是依赖关系,而 $AS=\{X\subseteq P(A)\mid X=\{\emptyset\}\vee\emptyset\notin X\}$;$I\in A\to AS$ 是输入绑定活动集;$O\in A\to AS$ 是输出绑定活动集;则

$D=\{(a_i,a_o)\in A\times A\mid a_1\in\bigcup_{as\in I(a_2)}as\}$

$D=\{(a_i,a_o)\in A\times A\mid a_2\in\bigcup_{as\in I(a_1)}as\}$

$\{a_i\}=\{a\in A\mid I(a)=\{\emptyset\}\}$

$\{a_o\}=\{a\in A\mid O(a)=\{\emptyset\}\}$

图(A,D)中的所有活动的路径是从 a_i 到 a_o。

定义 4[10](模式)　设 $C=(A,a_i,a_o,D,I,O)$是过程模型 M 的一个 C-网,A 的连通子图是 $P=(A',A_i',A_o',D',I',O')$,而 $A_i'\subseteq A'$,$A_o'\subseteq A'$分别表示开始活动和结束活动,P 是 M 的一个模式,当且仅当 $A'\subseteq A$;$D'\subseteq D$,对于任意的 $\alpha\in A'$,$I'(\alpha)\subseteq I(\alpha)$,$O'(\alpha)\subseteq O(\alpha)$。如果一个模式 $P=(A',A_i',A_o',D',I',O')$是一个简单模式,当且仅当对于所有的活动 $\alpha\in A'$:

$$[\exists!\Phi\in I'(\alpha):\Phi\not\subset R_\alpha^+]\vee[\forall\Phi\in I'(\alpha):\Phi\subseteq R_\alpha^+],$$

$$[\exists!\Theta\in O'(\alpha):\Theta\not\subset R_\alpha^-]\vee[\forall\Theta\in O'(\alpha):\Theta\subseteq R_\alpha^-]$$

其中,R_α^+ 是活动 α 的所有后集集合,R_α^- 是活动 α 的所有前集集合。

定义 5[14](k 阶继承关系)　设 $PN=(P,T;F)$ 是一个流程模型的 Petri 网,则 $\sigma\in PN$ 中的变迁对之间的 k 阶继承关系 $\triangleright_k^\sigma\subseteq T\times T$ 满足 $x\triangleright_k^\sigma y\Leftrightarrow\exists 1\leqslant i\leqslant|\sigma|[\sigma(i)=x\wedge\sigma(i+k)=y]$,其中,$k=1,2,3,\cdots$。网 N 的 k 阶继承关系为 $\triangleright_k^N\subseteq T\times T$,且满足 $x\triangleright_k^N y\Leftrightarrow\exists\sigma\in\sum[x\triangleright_k^\sigma y]$。变迁对之间的 k 阶继承关系为 $>_k^\sigma\subseteq T\times T$,且满足 $x>_k^\sigma y\Leftrightarrow\exists 1\leqslant 1\leqslant k,[x\triangleright_l^\sigma y]$。

当满足 $x\underline{\triangleright}_k^\sigma y\Leftrightarrow x\triangleright_k^\sigma y\wedge(x,y)\notin>_{k-1}^\sigma$时,则称其为最小 k 阶继承关系,记作 $\underline{\triangleright}_k^\sigma$。

4　基于 Petri 网行为紧密度的有效低频行为模式分析

本节主要介绍基于事件日志与流程模型间的行为紧密度发现有效的低频模式的方法,并提出了新的算法。该方法不仅解决了当前挖掘过程中只注重发现频繁行为而忽略低频行为模式的问题,还基于模式与流程模型间的行为关系检测发现了有效的低频行为模式,使其具备更好的适用性。首先,针对给定的事件日志找出合理的流程模型,基于流程模型测量其模式频率并将结果与阈值进行比较,从而发现其低频行为模式;通过判断低频模式中日志的行为关系是否符合模型的行为关系,并计算两者的行为紧密度,来检测有效的低频行为模式。

4.1　基于事件日志发现低频行为模式

过程挖掘主要是发现频繁行为,目的是了解过程中主要发生的事情,很少的工作涉及低频行为,并且已有工作主要集中在异常或偏差的检测上,忽略了低频行为对流程管理的重要性。本节提出了基于流程模型的低频模式的检测算法,算法 1 通过给定的事件日志找出合理的流程模型,利用阈值对流程模型中的初始模式进行迭代扩展,从而发现低频模式来研究低频行为。下面将给出 M-模式的定义,并基于活动的 M-模式进行迭代,从而找出低频行为模式。

首先,根据给定的事件日志建立流程模型,此时需要判断其合理性。根据文献[13],给出合理性的判断标准:

$$fitness=\frac{1}{2}\left(1-\frac{\sum_{i=1}^{k}n_im_i}{\sum_{i=1}^{k}n_ic_i}\right)+\frac{1}{2}\left(1-\frac{\sum_{i=1}^{k}n_ir_i}{\sum_{i=1}^{k}n_ip_i}\right)$$

其中,k 为给定日志的不同轨迹数,n 为日志轨迹中所含的数目,m 为日志轨迹中缺少的标识数,r 为日志轨迹中遗留的标识数,c 为日志轨迹中消耗的标识数,p 为日志轨迹中产生的标识数。

定义 6[10](最小模式,M-模式)　给定过程模型 M 的一个 C-网 $C=(A,a_i,a_o,D,I,O)$和一个活动 $\alpha'\in A$,如果一个模式 $P=(A',A_i',A_o',D',I',O')$是 α'的最小模式,当且仅当一个最大的简单模式包含 α',并且满足以下规则:

1)如果 $|I(\alpha')|>1$,那么$[I'(\alpha')=\emptyset]\vee[|I'(\alpha')|=1.\Phi\in I'(\alpha'):|\Phi|>1]$;

2)如果 $|O(\alpha')|>1$,那么$[O'(\alpha')=\emptyset]\vee[|O'(\alpha')|=1.\Theta\in O'(\alpha'):|\Theta|>1]$;

3)$\forall\alpha\in R_{\alpha'}^+$:如果$|O(\alpha)|\neq1$,那么 $O'(\alpha)=\emptyset$;

4)$\forall\alpha\in R_{\alpha'}^-$:如果$|I(\alpha)|\neq1$,那么 $I'(\alpha)=\emptyset$;

5)$\forall\alpha\in A',\alpha\neq\alpha',\alpha\in(R_{\alpha'}^+\cup R_{\alpha'}^-)$,如果$|I(\alpha)|\neq1$,那么 $I'(\alpha)=\emptyset$,并且如果$|O(\alpha)|\neq1$,那么 $O'(\alpha)=\emptyset$。

定义 7[10](低频模式)　设 L 是过程日志的迹集,一个简单模式 SP 的频率为 $freq(SP)=\frac{|\{\tau\in L:SP\vdash\tau\}|}{|L|}$,而模式 P 的频率为简单模式的最大频率,即:

$$freq(P)=\max_{\forall SP\in P}freq(SP)$$

给定一个频率阈值 $\sigma\in\mathcal{R}:0<\sigma\leqslant1$,一个模式 P 是一个低频模式当且仅当 $freq(P)<\sigma$。

算法 1　检测符合事件日志的流程模型的低频模式

输入:事件日志,合理性阈值 t_f,频率阈值 t_r

输出:符合事件日志的流程模型的低频模式

Step 1　根据给定的所有事件日志 L,分析各个变迁对 $t^1,t^2(t^1,t^2\in L)$的行为轮廓关系,建立初始的 Petri 网流程模型 M_0。计算

流程模型的合理性 fitness(M_0)，如果 fitness(M_0)$\geqslant t_f$，则模型合理，执行 Step 2；如果 fitness(M_0)$<t_f$，增加配置信息，调整流程模型，直至流程模型 M_i 满足 fitness(M_i)$\geqslant t_f$，执行 Step 2。

Step 2 按照实例数从大到小的顺序排列日志 L，如$\{\tau_1,\tau_2,\cdots,\tau_n\}$，计算日志的频率 freq(i)，i=1，2，…，n，如果 freq(i)$>t_r$，则将其删除，并构建剩余日志的流程模型；然后初始化流程模型中活动 α_i 的候选弧 $A^<$ 和 M-模式，对活动 α 的 M-模式进行扩展迭代，将初始迭代记为 P_0，扩展方式为：M← M∪addArcs($A^<$)或 M←M∪addMpattern(α_j)，这里 i≠j，执行 Step 3。

Step 3 根据定义 7，测量新模式的频率 freq(i)′，如果 freq(i)′>freq(i)，则将其舍弃；否则继续分析，执行 Step 4。

Step 4 若 freq(i)′≤freq(i)，则根据频率阈值 t_r 判断模式是否频繁，如果 freq(i)′$>t_r$，则模式是频繁模式，继续进行下一次迭代，重新执行 Step 2，否则执行 Step 5。

Step 5 如果 freq(i)′$<t_r$，且满足以下条件之一，则保留结果，执行 Step 6。

1)当前迭代为第一次迭代，即 P(i)=P_0；

2)迭代之前的频率大于阈值，迭代之后的频率小于阈值，即 freq(i)$>t_r$，freq(i)′$<t_r$；

3)迭代之后频率不变，即 freq(i)=freq(i)′。

Step 6 迭代完成，删除低频模式的冗余信息，输出流程模型的低频模式，算法结束。

4.2 基于行为紧密度的有效低频行为模式分析

通过以上研究发现的低频行为模式可能并不都是有效的低频模式，还可能存在噪音，因此需要过滤掉噪音，从而找出有效的低频行为模式。本节基于发现的低频模式提出了有效低频模式的分析算法，该算法根据行为紧密度的知识计算低频模式中的日志与模型间的行为紧密度，并给定行为紧密度的阈值，满足阈值条件的低频模式，即有效模式。下面首先给出了行为紧密度的定义。

定义 8[13]（日志距离向量） 设一条事件日志 $T=\{t_1,t_2,\cdots,t_n\}$，其中 $\forall(t_1,t_2)\in T$，$\forall(p_i,p_{i+1})\in t_i$ 的日志行为值为 $V_T(p_i,p_{i+1})$，且 $V_T(p_i,p_{i+1})=1$。日志的行为距离向量为：$X=\{V_T(p_1,p_2),V_T(p_2,p_3),\cdots,V_T(p_i,p_{i+1}),\cdots,V_T(p_j,p_m)\}=\{l_1,l_2,\cdots l_i,\cdots,l_m\}$。

类似地，可以定义模型的行为值为 $V_M(p_i,p_{i+1})=k$。模型的行为距离向量可定义为：$X=\{V_M(p_1,p_2),V_M(p_2,p_3),\cdots,V_M(p_i,p_{i+1}),\cdots,V_M(p_j,p_m)\}=\{k_1,k_2,\cdots,k_i,\cdots,k_m\}$。其中，$m$ 表示变迁的个数，且 $0<i\leqslant m$；k_i 为模型中与日志对应变迁的最小 k 阶继承关系。

定义 9[13]（行为紧密度） 给出事件日志 T 和流程模型 M，则日志 T 与模型 M 的行为紧密度定义如下：

$$\xi(T,M)=\frac{\sum_{i=1}^{m}l_i\times k_i}{\sqrt{(\sum_{i=1}^{m}l_i^2)\times(\sum_{i=1}^{m}k_i^2)}}$$

其中，m 表示日志中的活动行为关系数，k_i 为模型中与日志对应变迁的最小行为距离，l_i 为日志行为距离。

算法 2 基于行为紧密度的有效低频行为模式分析

输入：事件日志，合理性阈值 t_f，频率阈值 t_r，紧密度阈值 t_c

输出：有效低频行为模式

Step 1 根据算法 1，基于事件日志建立合理的流程模型，流程模型满足 fitness(M_i)$\geqslant t_f$；然后从流程模型中找出初始 M-模式进行迭代，以发现所有满足 freq(i)$<t_r$(i=1，2，…，n)的低频模式，执行 Step 2。

Step 2 整理从低频模式提取的低频日志 T_i，i=1，2，…，m，求出模型与日志对应变迁的最小 k 阶继承关系 $\underline{\triangleright}_k^\sigma$，即 $x\ \underline{\triangleright}_k^\sigma y\Leftrightarrow x\triangleright_k^\sigma y\wedge(x,y)\notin>_{k-1}^\sigma$。求出日志的行为距离向量$\{l_1,l_2,\cdots,l_i,\cdots,l_m\}$和模型的行为距离向量$\{k_1,k_2,\cdots,k_i,\cdots,k_m\}$，执行 Step 3。

Step 3 根据定义 9，根据行为紧密度公式计算出每个低频日志 T_i 与合理流程模型 M_i 的行为紧密度 $\xi(T_i,M)$，执行 Step 4。

Step 4 如果日志与流程模型的行为紧密度值大于或等于给定的紧密度阈值 t_c，即 $\xi(T_i,M)\geqslant t_c$，则该低频日志为有效低频日志，将其保留下来；如果 $\xi(T_i,M)<t_c$，则将该低频日志视为噪音，直接删除，执行 Step 5。

Step 5 输出所有保留的事件日志的模式，即有效低频行为模式，算法结束。

5 实例分析

本节以上述车辆保险理赔系统（动机例子）为例来验证所提算法的可行性。给定合理性阈值 $fitness(M)=1$，频率阈值为 10%（因判断模型中的低频模式，模拟系统的不常见行为，且给定事件的日志实例数较少，故给定阈值为 10%），行为紧密度阈值为 90%（因事件日志与模型存在行为差异，日志与模型的行为距离向量不一致，故给定行为紧密度阈值为 90%）。第 2 节已经建立了一个初始的流程模型，下面通过合理性阈值来判断模型的合理性，并发现流程模型中的低频行为模式，计算从低频模式中提取的日志与模型的紧密度值，从而找出有效的低频行为模式。

首先，观察和计算如图 1 所示的车辆保险理赔的初始模型 M_0 的合理性 $fitness(M_0)=0.9868\neq1$。可知初始模型 M_0 不完全合理，因此首先根据表 2 的事件日志的行为轮廓关系建立行为轮廓关系表，如表 3 所列；然后增加配置[15]信息，将模型调整为带配置信息的 Petri 网模型 M_1（见图 2）；最后计算出合理性值 $fitness(M_1)=1$，满足要求。

将事件日志按照实例数从多到少的顺序进行排列，结果如下：

{〈ABCDEFGKLMNOPQ〉^{2904}，〈ABCDEFGHJLMNOPQ〉^{2556}，〈ABCEDFGKLMNMNOPQ〉^{2288}，〈ABCEDFGHJLMNMNOPQ〉^{734}，〈ABCEDFGHILMNMNOPQ〉^{695}，〈ABCDEFGKLMNRQ〉^{648}，〈ABCDEFGHJLMNRQ〉^{43}，〈ABCEDFGHILMNRQ〉^{35}，〈ABCDEFGKLRQ〉^{28}，〈ABCEDFGHJLRQ〉^{23}，〈ABCDEFQ〉^{18}，〈ABCDEFGHILRQ〉^{15}，〈ABCDEFGHILRQ〉^{13}}

为了发现低频的行为模式，需要先把频数较低的日志挖掘出来。根据上述事件日志实例数的排列情况，将案例 5—案例 7 的事件日志删除，并对剩余日志中活动的 M-模式初始化后开始迭代。根据计算可知，上述 Petri 网模型 M_1 中的几

个 M-模式出现的频数较低（见图 3），因此分别以它们作为初始模式开始迭代。根据算法 1 可知，模式 1 和模式 2 在第一次迭代完成后就已经是低频模式，其中 $freq(a)=0.0015<0.1$，$freq(b)=0.0054<0.1$，而模式 3 需要进行迭代，当扩展迭代为迹是〈L，M，N，R，Q〉的模式时，即为低频模式，频率为 0.0106，如图 4 所示。从图 4 中可知，迭代完成后的模式 3 包含了模式 2，因此模式 2 可直接省略；其余模式因为频率大于给定的频率阈值，所以直接删除。

表 3　行为轮廓关系表

Table 3　Behavioral profiles relation table

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
A	+	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→
B		+	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→
C			+	→	→	→	→	→	→	→	→	→	→	→	→	→	→	→
D				+	‖	→	→	→	→	→	→	→	→	→	→	→	→	→
E					+	→	→	→	→	→	→	→	→	→	→	→	→	→
F						+	→	→	→	→	→	→	→	→	→	→	→	→
G							+	→	→	→	→	→	→	→	→	→	→	→
H								+	→	→	+	→	→	→	→	→	→	→
I									+	+	+	→	→	→	→	→	→	→
J										+	+	→	→	→	→	→	→	→
K											+	→	→	→	→	→	→	→
L												+	→	→	→	→	→	→
M													‖	‖	→	→	→	+
N														‖	→	→	→	+
O																→	→	+
P																	→	+
Q																	+	
R																→		

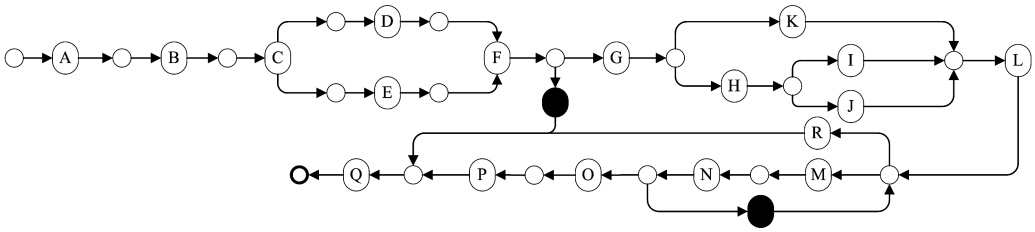
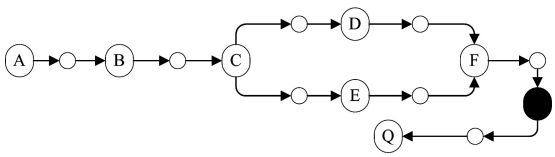


图 2　车辆保险理赔的 Petri 网模型 M_1

Fig. 2　Petri net model M_1 of vehicle insurance claims



(a)模式 1



(b)模式 2　　(c)模式 3

图 3　频率较低的几个模式

Fig. 3　Several patterns with low-frequency

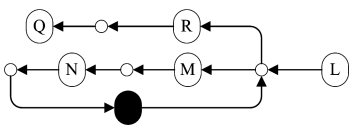


图 4　迭代完成的模式 3

Fig. 4　Iteration completed pattern 3

下面分别计算模型 M_1 与两个低频模式提取的事件日志的行为紧密度。首先提取模式中的事件日志：模式 1 为〈A，B，C，D，E，F，Q〉，模式 3 为〈L，M，N，R，Q〉或〈L，R，Q〉。日志 ABCDEFQ 的行为距离向量 $X_1=(1,1,1,1,1,1)$，日志 LMNRQ 的行为距离向量 $X_2=(1,1,1,1)$，日志 LRQ 的行为距离向量 $X_3=(1,1)$。模型 M_1 对应变迁的最小行为距离向量 $X_1'=(1,1,1,1,1,2)$，$X_2'=(1,1,2,1)$，$X_3'=(1,1)$。根据行为紧密度公式可得：

$$\xi_1=\frac{1\times1+1\times1+1\times1+1\times1+1\times1+1\times1+1\times2}{\sqrt{1^2+1^2+1^2+1^2+1^2+1^2}\times\sqrt{1^2+1^2+1^2+1^2+1^2+2^2}}\approx0.9526$$

$$\xi_2=\frac{1\times1+1\times1+1\times2+1\times1}{\sqrt{1^2+1^2+1^2+1^2}\times\sqrt{1^2+1^2+2^2+1^2}}\approx0.9449$$

$$\xi_3=\frac{1\times1+1\times1}{\sqrt{1^2+1^2}\times\sqrt{1^2+1^2}}=1$$

经计算可得，3 个日志与模型的行为紧密度都大于阈值 0.9，因此 3 个低频行为模式都是有效的。将低频模式放到本节实例中，不难知道，一般车主在为车辆入保后，需要严格按照合同内容执行，并完整提交理赔所需的材料，这样就能避免车辆在出现事故以后无法受保的情况。同时，针对少数可能存在材料不完善或者出现欺诈性索赔的行为，保险公司需要完善理赔程序。

结束语 本文在已有研究的基础上,基于日志与模型的行为紧密度对低频行为模式进行分析,并找出有效的低频行为模式。本文研究表明了过程挖掘不仅要区分频繁行为与低频行为,还要分析低频行为是有效低频还是噪音。首先,根据日志建立合理的流程模型,找出低于阈值的低频行为模式,利用算法进行迭代扩展可以极大地减少人工计算量;其次,计算日志与模型的行为距离向量,从而求出行为紧密度,以此找出符合模型的有效低频行为模式。利用所提方法发现的低频行为模式能应用于实际生活,弥补了很多系统的不足之处。

关于低频行为的发现还有很广阔的研究空间,提出的方法只适用于简单、合理的流程模型,对于复杂的流程模型需进一步研究;流程模型中的有效低频行为和噪音也是未来的研究方向之一。

参 考 文 献

[1] VÁZQUEZ-BARREIROS B, MUCIENTES M, LAMA M. ProDiGen: Mining complete, precise and minimal structure process models with a genetic algorithm[J]. Information Sciences, 2015, 294: 315-333.

[2] LEEMANS S J J, FAHLAND D, AALST W M P V D. Discovering Block-Structured Process Models from Event Logs-A Constructive Approach[M]// Application and Theory of Petri Nets and Concurrency. Springer Berlin Heidelberg, 2013: 311-329.

[3] TAX N, SIDOROVA N, HAAKMA R, et al. Mining local process models[J]. Journal of Innovation in Digital Ecosystems, 2016, 3(2): 183-196.

[4] SURIADI S, MANS R S, WYNN M T, et al. Measuring Patient Flow Variations: A Cross-Organisational Process Mining Approach[M]// Asia Pacific Business Process Management. Springer International Publishing, 2014: 43-58.

[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] MANNHARDT F, LEONI M D, REIJERS H A, et al. Data-Driven Process Discovery-Revealing Conditional Infrequent Behavior from Event Logs[C]// International Conference on Advanced Information Systems Engineering. Springer, Cham, 2017: 545-560.

[7] LIESAPUTRA V, YONGCHAREON S, CHAISIRI S. Efficient Process Model Discovery Using Maximal Pattern Mining[C]// International Conference on Business Process Management. Springer, Cham, 2015: 441-456.

[8] LU X, FAHLAND D, BIGGELAAR F J H M V D, et al. Detecting Deviating Behaviors Without Models[C]// International Conference on Business Process Management. Springer, Cham, 2015: 126-139.

[9] LEEMANS S J J, FAHLAND D, AALST W M P V D. Discovering Block-Structured Process Models from Event Logs Containing Infrequent Behaviour[C]// International Conference on Business Process Management. Springer, Cham, 2013: 66-78.

[10] 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.

[11] JENSEN M T. Improving robustness and flexibility of tardiness and total flow-time job shops using robustness measures[J]. Applied Soft Computing, 2001, 1(1): 35-52.

[12] WU Z H. Petri nets theory [M]. Beijing: Mechanical Industry Press, 2006: 6-22. (in Chinese)
吴哲辉. Petri 网理论[M]. 北京:机械工业出版社, 2006: 6-22.

[13] GAO Y N, FANG X W, WANG L L. Optimized analysis of business process configuration based on Petri net behavior closeness [J]. Computer Science, 2017, 44(S1): 539-542. (in Chinese)
高雅楠,方贤文,王丽丽. 基于 Petri 网行为紧密度的业务流程配置优化分析[J]. 计算机科学, 2017, 44(S1): 539-542.

[14] CHEN L N, FANG X W, LIU X W. Optimized analysis of business process model query of Petri net behavior inclusion[J]. Computer Engineering and Design, 2017, 38(2): 409-413. (in Chinese)
陈丽娜,方贤文,刘祥伟. Petri 网行为包含的业务流程模型查询优化分析[J]. 计算机工程与设计, 2017, 38(2): 409-413.

[15] AALST W M P V D, DUMAS M, GOTTSCHALK F, et al. Preserving correctness during business process model configuration [J]. Formal Aspects of Computing, 2010, 22(3-4): 459-482.