# 基于通讯行为轮廓挖掘条件非频繁行为的方法

曹 蕊 方贤文 王丽丽

(安徽理工大学数学与大数据学院 安徽 淮南 232001)

**摘 要** 条件非频繁行为是指带有属性值的频数较低事件轨迹所记录的行为。从记录的事件日志中挖掘条件非频繁行为是业务过程优化的主要内容之一。已有的方法删除低频次行为,较少考虑模块网间数据流角度下的条件非频繁行为。基于此,文中提出了基于通讯行为轮廓挖掘条件非频繁行为的方法。以模块网间的通讯行为轮廓理论为基础,首先,通过给定的业务过程源模型查找其可执行事件日志,并且找出频数较低的事件轨迹,添加相关属性及属性值,即可得到条件非频繁轨迹;其次,通过计算不同模块网间通讯特征的条件依赖数值,确定条件不频繁轨迹是否删除或保留,从而得到优化事件日志,进而挖掘出优化通讯模型;最后,通过仿真实验验证了该方法的可行性。

**关键词** Petri网,通讯行为轮廓,业务过程模型,非频繁行为,模块网,噪音

**中图法分类号** TP391.9 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.08.056

## Method of Mining Conditional Infrequent Behavior Based on Communication Behavior Profile

CAO Rui FANG Xian-wen WANG Li-li

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

**Abstract** Conditional infrequent behavior refers to the behavior recorded by infrequent event traces with attribute values. Mining the conditional infrequent behavior from the event log is one of the main contents of business process optimization. The existing methods remove low frequency behavior, but take less consideration of the conditional infrequent behavior under the perspective of data-flow between different module nets. Based on this, the paper presented a method of mining conditional infrequent behavior based on communication behavior profile. Based on the communication behavior profile theory between module nets, firstly, through a given business process source model, its executable event log is searched and the infrequent event traces are found, adding the relevant attributes and attribute values to the infrequent event traces to get the conditional infrequent traces. Secondly, by calculating condition dependent values of the communication features of different module nets, whether the conditional infrequent traces are deleted or retained can be determined. The optimized event log is given, and the business process optimization communication model is mined. Finally, the feasibility of the method is verified by a simulation.

**Keywords** Petri net, Communication behavioral profile, Business process model, Infrequent behavior, Module net, Noise

## 1 引言

在实际生活中,从信息系统中提取的事件日志往往包含了被称为噪音的条件非频繁行为。在业务过程执行中,虽然它们出现的频次相对较少,但有些条件非频繁行为却占据着重要的地位,是业务流程主要行为的一个分支,不可轻易地全部过滤掉。例如,在医疗机构系统中,由于意外情况出现的急诊行为是不常见行为,但它却是医疗业务过程中的重要行为之一。因此,挖掘主要的非频繁行为具有一定的意义。

目前,针对非频繁行为方面的相关研究工作已较多。文献[1-2]指出了在流程挖掘中处理包含噪音的事件日志是一个重大挑战。文献[3]提出了事件日志中总是包含噪音等的非频繁行为(如无序事件、记录错误或者异常行为),且研究了支持噪音过滤的流程发现方法。文献[4-7]提出了基于频率的噪音过滤方法、机器学习技术、概率模型、专用噪音过滤方法。文献[8]提出了一种以域理论为基础,基于整数线性规划的流程发现方法,此方法能够挖掘出一个松弛合理的工作流网。文献[9]研究了过程发现,旨在从给定事件日志中更好地获取描述流程模型的行为,然而非频繁行为的存在影响了获取模型的精确性。本文通过 Inductive Miner,使用切过滤非

到稿日期:2017-07-24 返修日期:2017-11-19 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),安徽省学术和技术带头人资助项目(DG119),安徽省优秀青年人才项目(ZY290)资助。

曹 蕊(1990-),女,硕士生,主要研究领域为 Petri 网,E-mail:734526382@qq.com;方贤文(1975-),男,博士,教授,主要研究领域为 Petri 网和可信软件,E-mail:280060673@qq.com(通信作者);王丽丽(1982-),女,副教授,主要研究领域为业务流程分析和软件认证。

频繁行为,挖掘得到合理的过程模型,并从质量和性能方面将该方法与现有的挖掘方法进行对比,从而证明了其有效性。

本文以通讯行为轮廓理论为基础,给定业务过程源模型,查询其可执行事件日志,并找出频数较低的事件轨迹,添加属性值后对其进行属性化。求出频数较低事件轨迹中具有直接依赖关系的事件对的数据依赖关系值,从而确立条件非频繁事件轨迹是否被删除,进而得到优化事件的轨迹集,并发现业务过程优化通讯模型。

本文第 2 节介绍准备知识;第 3 节提出基于通讯行为轮廓挖掘条件非频繁行为的方法,其中 3.1 节介绍业务过程条件非频繁行为,3.2 节提出基于通讯行为轮廓挖掘条件不频繁行为的算法;第 4 节通过仿真实验验证所提方法的可行性;最后总结全文并展望未来。

## 2　准备知识

通讯行为轮廓以通讯后继关系为基础。

**定义 1**[10](通讯后继关系)　设 $L\subseteq T^*$ 是事件日志,$T$ 是 Petri 网中的变迁集,通过 $A\prec_L B$ 定义通讯后继关系 $\prec_L\subseteq T\times T$ 当且仅当 $\Re(A)\neq\Re(B)$,即 $A,B$ 属于不同的模块,$\sigma(i)=A,\sigma(i+1)=B$。其中,事件轨迹 $\sigma\in L,1\leqslant i<|\sigma|$。

**定义 2**[10](通讯行为轮廓)　设 $L\subseteq T^*$ 是事件日志,$\prec_L\subseteq T\times T$ 是相应的通讯后继关系。通讯行为轮廓是一个三元数组 $(\rightarrow_C,\|_C,+_C)^{Com}$,它由以下关系组成:

1)严格通讯关系 $A\rightarrow_C B$,当且仅当 $A\prec_L B,B\not\prec_L A$;

2)交叉通讯关系 $A\|_C B$,当且仅当 $A\prec_L B,B\prec_L A$;

3)排它通讯关系 $A+_C B$,当且仅当 $A\not\prec_L B,B\not\prec_L A$;

4)逆严格通讯关系 $A\leftarrow_C B$,当且仅当 $A\not\prec_L B,B\prec_L A$。

**定义 3**[10](特征网)　设 $L\subseteq T^*$ 是一个事件日志,$A,F\in T$ 是特征(变迁)。设 $(\rightarrow_C,\|_C,+_C)^{Com}$ 是相应的通讯行为轮廓,特征网 $N_F$ 满足以下条件:

1)$P=\overline{P},T=\overline{T},i=[\bar{i}],\Omega=\{[\bar{f}]\}$;

2)$I=\{p_{A-F}|A\rightarrow F\}$;

3)$O=\{p_{F-A}|F\rightarrow A\}$;

4)$F=\overline{F}\cup\{(t,p_{F-A})|t\in T,\lambda(t)=A,F\rightarrow A\}\cup\{(p_{F-A},t)|t\in T,\lambda(t)=A,A\rightarrow F\}$。

其中,$I$ 和 $O$ 是端口库所,$\langle\overline{P},\overline{T},\overline{F},\bar{i},\bar{f}\rangle$ 是工作流网。工作流网 $WFN=\langle N,i,f\rangle$,其中 $N=\langle P,T,F\rangle$ 为 Petri 网,满足如下条件:1)$i\in P$ 且无入弧;2)$f\in P$ 且无出弧;3)所有的变迁至少有一个入弧和一个出弧,即 $\forall\tau\in T,{}^\cdot\tau\neq\emptyset\neq\tau^\cdot$。另外,本文中的模块网是 Petri 网,并且作了标注。

给定属性 $A$ 的全集和值 $U$ 的全集,$E$ 是有限事件标识符集,映射 $\sharp:E\rightarrow(A\not\rightarrow U)$ 为事件记录的属性值,$val:E\rightarrow(A\not\rightarrow U)$ 是指一个事件发生前所记录的最新属性值,即 $val(e_i)=val(e_{i-1})\oplus\sharp(e_{i-1})$,其特殊形式为 $val(e_1)=f_\emptyset$,$f\oplus g$ 是指 $f$ 和 $g$ 的复合覆盖,$f_\emptyset:\emptyset\rightarrow U$ 是一个空映射。

**定义 4**[11](依赖条件)　给定属性全集 $A$、值的全集 $U$ 和活动集 $\Sigma\subseteq U$,定义依赖条件为:

$C\in(\Sigma\times\Sigma)\rightarrow((A\not\rightarrow U)\rightarrow\{0,1\})$

对于属性值 $x\in(A\not\rightarrow U)$,依赖条件 $C_{a,b}(x)=(C(a,b))(x)$ 是指预知活动 $b$ 是否直接跟随着活动 $a$ 的二元分类器,即当 $b$ 直接跟随着 $a$ 时,$C_{a,b}(x)=1$,否则 $C_{a,b}(x)=0$。

**定义 5**[11](条件直接跟随关系)　给定活动 $a,b\in\Sigma$ 和依赖条件 $C$,在依赖条件 $C_{a,b}(x)$ 下,满足最新的属性值 $x$,称 $a>^{C,L}b$ 当且仅当活动 $b$ 的执行直接跟随着活动 $a$ 的发生。在事件日志中,$\perp$ 是空事件,$abs$ 表示集合中 $e$ 出现的个数,条件直接跟随关系 $a>^{C,L}b$ 的出现频数的定义如下:

$$|a>^{C,L}b|=abs\{e\in E|\sharp_{act}({}^\cdot(e))=a\wedge{}^\cdot(e)\neq\perp\wedge\sharp_{act}(e)=b\wedge C_{a,b}(val(e))=1\}$$

有关条件依赖度量的定义及其公式请参考文献[11]中的定义 3,这里不再赘述。

**定义 6**[11](条件度量方法)　设 $(\rightarrow_C,\|_C,+_C)^{Com}$ 是通讯行为轮廓,给定不同的活动 $a,b\in\Sigma$ 和依赖条件 $C$,满足依赖条件 $C_{a,b}(x)=1$ 下,即活动 $b$ 的发生直接跟随着活动 $a$ 的发生,定义:$a\Rightarrow^{C,L}b:\Sigma\times\Sigma\rightarrow[-1,1]$ 为从 $a$ 到 $b$ 的因果依赖的强度。其中:

$$a\Rightarrow^{C,L}b=\begin{cases}\dfrac{|a\rightarrow_C b|-|b\rightarrow_C a|}{|a\rightarrow_C b|+|b\rightarrow_C a|+0.1}, & a\rightarrow_C b\\ \dfrac{|a+_C b|-|b+_C a|}{2(|a+_C b|+|b+_C a|)+0.1}, & a+_C b\\ \dfrac{|a\|_C b|+|b\|_C a|}{|a\|_C b|+|b\|_C a|+0.1}, & a\|_C b\end{cases}\tag{1}$$

## 3　基于通讯行为轮廓挖掘条件非频繁行为的方法

条件非频繁行为是指属性化的频数较低的事件轨迹所记录的行为。在操作过程中,这些条件非频繁行为往往被直接删除,使得业务过程模型遗失了重要的行为,不能充分达到需求者的要求,因此需要对非频繁行为实行择优保留的方案进行分类处理。

对条件非频繁行为进行处理,已有的研究方法大多是忽略低频次行为,很少考虑事件间的数据依赖关系,尤其是模块间通讯事件(特征)的数据依赖关系。本文以模块网之间的通讯行为轮廓关系为基础,计算通讯特征(事件)条件依赖数据关系,对条件非频繁行为进行分类优化处理。

### 3.1　业务过程条件非频繁行为

在过程挖掘中,许多算法的提出往往是以事件日志能够记录业务过程的完整规则行为为基础的,进而来挖掘业务过程模型。然而,在实际生活中的事件日志总是包含各种各样的异常(非频繁)行为,而这些行为通常被认为是噪音,噪音的存在使得发现的模型变得繁琐,不能正常表示出实际的行为,甚至使业务流程模型不能有效运行。

对频数较低的事件日志轨迹添加了相关属性(如标识符、活动等)以及规则约束条件,称之为条件非频繁轨迹,其记录了条件非频繁行为。

### 3.2　条件非频繁行为的挖掘算法

条件非频繁行为是业务流程模型记录行为中不太常见的行为,在大多数情况下将其统称为噪音并删除,致使某些系统

模型遗失了部分重要信息，严重地影响了企业或者组织的操作效率。为了弥补此缺陷，对非频繁行为实行分类处理（并非全部删除）变得尤其重要。

对于频数较低的可执行迹（由 $t_{1,i}$ 和 $t_{2,j}$ 按照一定的发生顺序构成，其中，$t_{1,i}$ 和 $t_{2,j}$（$1\leqslant i\leqslant n, 1\leqslant j\leqslant n, n\in N_+$）是指两个模块网中不同模块的活动或变迁），添加相关的属性及其对应值，得到条件非频繁轨迹。本文方法基于行为轮廓理论计算具有直接跟随关系的活动（特征）对的条件依赖度量值，并以此判断条件不频繁行为是否为噪音，进而确定条件非频繁轨迹是否保留。源模型的可执行轨迹集包括频繁轨迹和非频繁轨迹。本文的频繁轨迹是指无记录错误，且能够正确拟合参考模型行为的轨迹。具体的方法如算法 1 所示。

**算法 1** 基于通讯行为轮廓挖掘条件不频繁行为的算法

输入：源模型 $M_O$，阈值 $\theta$

输出：业务过程优化通讯模型 $M_T$

1. 遍历输入的 Petri 网业务流程源模型 $M_O$，查询源模型 $M_O$ 中的可执行轨迹，并舍去不完备的事件轨迹，得到的可执行轨迹记为 $\sigma_1$，$\sigma_2$，…，$\sigma_n$，可执行事件日志记作 $L=\{\sigma_1,\sigma_2,\cdots,\sigma_n\}$，$n=1,2,3,\cdots$。查找频数较低的事件轨迹 $\sigma_1,\sigma_2,\cdots,\sigma_k$，$k<n$，$k\in N_+$。
2. 对需进行预处理的频数较低的事件轨迹（$\sigma_1,\sigma_2,\cdots,\sigma_k$，$k<n$，$k\in N_+$）添加相关的属性值，进行属性化操作，并满足一定的规则约束条件。
3. 根据步骤 2 中属性化的条件非频繁可行迹，在规则约束条件下，通过定义 4 中的二元分类器揭示活动（事件或特征）对间的条件直接跟随关系，依据定义 5 计算条件直接跟随关系 $a>^{C,L}b$ 的出现频数。
4. 根据条件依赖度量的公式[11]计算从活动 a 到活动 b 的因果依赖的强度 $a\Rightarrow^{C,L}b$ 的绝对值。
5. 在规则约束条件下，根据步骤 2 中属性化的条件非频繁可执行轨迹集、可执行轨迹集（$L=\{\sigma_1,\sigma_2,\cdots,\sigma_n\}$，$n=1,2,3,\cdots$）中的频繁轨迹集和定义 2 中的通讯行为轮廓，查询属于不同模块网中活动（事件或特征）对的频数较低轨迹通讯行为的轮廓关系，并计算其权值。例如，$|a\rightarrow_C b|=|\{a\rightarrow_C b\mid a\prec b\wedge b\not\prec a\wedge \#_{act}(\cdot(e))=a\wedge \cdot(e)\neq\perp\wedge\#_{act}(e)=b\wedge\mathscr{R}(a)\neq\mathscr{R}(b)\wedge C_{a,b}(val(e))=1\}|$，权值 $|a+_C b|$，$|a||_C b|$ 及 $|b\rightarrow_C a|$，$|b+_C a|$，$|b\|_C a|$ 的计算同上。
6. 根据定义 6 中的式（1）计算从活动 a 到活动 b 的因果依赖的强度 $a\Rightarrow^{C,L}b$ 的绝对值。
7. 以引入的阈值 $\theta$ 为基准，对步骤 4 的计算结果与步骤 6 的计算结果进行比较和分析。
8. 如果 $0<|a\Rightarrow^{C,L}b|<\theta$，表明从 a 到 b 的因果依赖的强度较弱，此类条件非频繁行为为噪音，去除包含此噪音的事件轨迹。
9. 如果 $|a\Rightarrow^{C,L}b|\geqslant\theta$，说明从 a 到 b 的因果依赖的强度较强，包含此类条件非频繁行为的事件轨迹即非噪音的轨迹被保留，用来发现优化模型，记为 $\sigma_1,\sigma_2,\cdots,\sigma_p$，包含非噪音的频数较低的事件日志轨迹集记为 $L_{con}=\{\sigma_1,\sigma_2,\cdots,\sigma_p\}$，$p\leqslant k$，$p\in N_+$。
10. 根据包含非噪音频数较低事件轨迹集（$L_{con}=\{\sigma_1,\sigma_2,\cdots,\sigma_p\}$，$p\leqslant k$，$p\in N_+$）的优化事件轨迹集 $L_N=\{\sigma_1,\sigma_2,\cdots,\sigma_q\}$，$q\leqslant n$，$q\in N_+$，利用 $\alpha$ 算法[12]构建业务过程优化通讯模型 $M_T$。
11. 输出业务过程优化通讯模型 $M_T$。

## 4 仿真实验

为了验证算法 1 的可行性，将对门诊患者的就诊业务过程进行分析。首先，患者进入医院挂号，购买门诊病历并填写完整，然后等待医生接诊。接诊后，依据医生的意见，检查身体或者取药治疗，最后根据病情确定住院治疗或者回家休养等治疗方式。具体的源模型 $M_O$ 如图 1 所示。其中，$t_{11}$ 为挂号，$t_{12}$ 为医生接诊，$t_{13}$ 为检查，$t_{14}$ 为医生给处方，$t_{15}$ 为回家休养，$t_{16}$ 为住院治疗，$t_{17}$ 为复诊，$t_{18}$ 为离院，$t_{1i}$（$1\leqslant i\leqslant 8$）是源模型 $M_O$ 中模块网 $M'$ 的变迁。$t_{21}$ 为填写门诊病历，$t_{22}$ 为护士分诊，$t_{23}$ 为候诊，$t_{24}$ 为优先接诊，$t_{25}$ 为拿到检查结果，$t_{26}$ 为取药治疗，$t_{17}$ 为病情稍重，$t_{18}$ 为病情轻微，$t_{2j}$（$1\leqslant j\leqslant 8$）是源模型 $M_O$ 中模块网 $M''$ 的变迁。

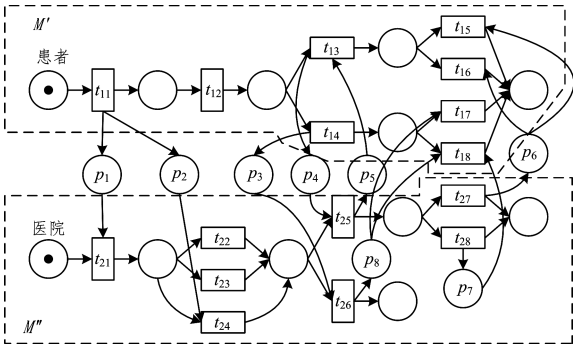


图 1 Petri 网业务过程源模型 $M_O$

Fig. 1 Business process source model $M_O$ of Petri net

通过图 1 中的业务流过程源模型 $M_O$，可以找出其执行的完备的可执行迹（如表 1 所列，此处的数据为实验平台设计数据），包括频数较低的可行迹 $\sigma_6$，$\sigma_7$，$\sigma_8$，$\sigma_9$。对频数较低的轨迹添加相关属性：标识符，活动，是否检查，治疗类别。具体的带有属性值的频数较低的可行迹如表 2 所列。给定规则约束条件：患者挂号后优先接诊，进行体检得到结果后，如果病情稍重，则选择住院治疗（需满足时间的先后顺序）。

表 1 可执行事件日志

Table 1 Executable event log

| 编号 | 频数 | 可执行事件日志轨迹 |
|---|---|---|
| 1 | 556 | $t_{11}t_{21}t_{23}t_{12}t_{13}t_{25}t_{28}t_{18}$ |
| 2 | 997 | $t_{11}t_{21}t_{22}t_{12}t_{14}t_{26}$ |
| 3 | 679 | $t_{11}t_{21}t_{12}t_{22}t_{13}t_{25}t_{27}t_{15}$ |
| 4 | 778 | $t_{11}t_{21}t_{23}t_{12}t_{14}t_{26}t_{17}$ |
| 5 | 989 | $t_{11}t_{21}t_{12}t_{22}t_{14}t_{26}t_{18}$ |
| 6 | 23 | $t_{11}t_{21}t_{12}t_{24}t_{13}t_{25}t_{27}t_{16}$ |
| 7 | 7 | $t_{21}t_{11}t_{24}t_{12}t_{25}t_{27}t_{16}t_{13}$ |
| 8 | 11 | $t_{11}t_{21}t_{27}t_{25}t_{12}t_{24}t_{13}t_{16}$ |
| 9 | 2 | $t_{24}t_{12}t_{25}t_{13}t_{21}t_{11}t_{16}t_{27}$ |
| … | … | … |

根据算法 1 中的步骤 3 和步骤 4，依据满足规则约束条件非频繁轨迹集，计算不同模块间具有条件直接跟随关系的活动对的出现频数，例如，$|t_{11}>^{C,L}t_{21}|=23+11=34$，$|t_{12}>^{C,L}t_{25}|=7+2=9$。利用文献[11]中的方法计算条件依赖度量的绝对值，例如，$|t_{11}\Rightarrow^{C,L}t_{21}|=\left|\frac{23+11-7-2}{23+11+7+2+1}\right|\approx 0.568$，$|t_{12}\Rightarrow^{C,L}t_{25}|=\left|\frac{7+2-11}{7+2+11+1}\right|\approx 0.095$。

表 2　带有属性(标识符、活动、是否检查、治疗类别)的 4 条轨迹

Table 2　Four traces of example process with attributes identifier, activity, check or no check, treatment category

迹 $\sigma_6 \in L$

| 标识符 | 活动 | 是否检查 | 治疗类别 |
|---|---|---|---|
| $t_{11}$ | 挂号 | | |
| $t_{21}$ | 填病历 | | |
| $t_{12}$ | 医师接诊 | | |
| $t_{24}$ | 优先接诊 | | |
| $t_{13}$ | 检查 | 是 | |
| $t_{25}$ | 检查结果 | | |
| $t_{27}$ | 病情稍重 | | |
| $t_{16}$ | 住院 | | 住院 |

迹 $\sigma_7 \in L$

| 标识符 | 活动 | 是否检查 | 治疗类别 |
|---|---|---|---|
| $t_{21}$ | 填病历 | | |
| $t_{11}$ | 挂号 | | |
| $t_{24}$ | 优先接诊 | | |
| $t_{12}$ | 医师接诊 | | |
| $t_{25}$ | 检查结果 | | |
| $t_{27}$ | 病情稍重 | | |
| $t_{16}$ | 住院 | | 住院 |
| $t_{13}$ | 检查 | 是 | |

迹 $\sigma_8 \in L$

| 标识符 | 活动 | 是否检查 | 治疗类别 |
|---|---|---|---|
| $t_{11}$ | 填病历 | | |
| $t_{21}$ | 病情稍重 | | |
| $t_{27}$ | 检查结果 | | |
| $t_{25}$ | 医师接诊 | | |
| $t_{12}$ | 挂号 | | |
| $t_{24}$ | 优先接诊 | | |
| $t_{13}$ | 检查 | 是 | |
| $t_{16}$ | 住院 | | 住院 |

迹 $\sigma_9 \in L$

| 标识符 | 活动 | 是否检查 | 治疗类别 |
|---|---|---|---|
| $t_{24}$ | 优先接诊 | | |
| $t_{12}$ | 医师接诊 | | |
| $t_{25}$ | 检查结果 | | |
| $t_{13}$ | 检查 | 是 | |
| $t_{21}$ | 填病历 | | |
| $t_{11}$ | 挂号 | | |
| $t_{16}$ | 住院 | | 住院 |
| $t_{27}$ | 病情稍重 | | |

根据算法 1 中的步骤 5 和步骤 6，依据源模型可执行迹中的频繁轨迹和满足规则约束条件的频次较低的轨迹集，查找频数较低事件轨迹中属于不同模块活动(事件或特征)对的通讯行为轮廓关系(见表 3)，并计算其权值。例如，变迁(事件或特征)对的权值为 $|t_{11} \rightarrow_C t_{21}| = 23+11 = 34$，$|t_{12} +_C t_{25}| = 9$。根据定义 6 中的式(1)计算从活动 $a$ 到活动 $b$ 的因果依赖的强度 $a \Rightarrow^{C,L} b$ 的绝对值，例如：

$$|t_{11} \Rightarrow^{C,L} t_{21}| = \left|\frac{23+11-7-2}{23+11+7+2+0.1}\right| \approx 0.575$$

$$|t_{12} \Rightarrow^{C,L} t_{25}| = \left|\frac{7+2-11}{2(7+2+11)+0.1}\right| \approx 0.050$$

表 3　通讯行为轮廓

Table 3　Communication behavior profile

| | $t_{21}$ | $t_{24}$ | $t_{25}$ | $t_{27}$ |
|---|---|---|---|---|
| $t_{11}$ | $\rightarrow_C$ | $+_C$ | $+_C$ | $+_C$ |
| $t_{12}$ | $\|_C$ | $\|_C$ | $+_C$ | $+_C$ |
| $t_{13}$ | $\|_C$ | $\|_C$ | $\rightarrow_C$ | $+_C$ |
| $t_{16}$ | $\|_C$ | $+_C$ | $\|_C$ | $\leftarrow_C$ |

分别采用文献[11]中的方法和本文提出的方法计算活动(事件)对的条件依赖度量的绝对值，结果如表 4 所列。引入阈值 $\theta = 0.500$，为了更加直观形象，基于表 4 中的结果数据建立折线图，如图 2 所示。

表 4　不同方法针对直接跟随关系活动对的结果比较

Table 4　Result comparison of different methods for activity pairs with direct following relationship

| 编码 | 度量绝对值表示 | 文献[11]中的方法 | 本文方法 |
|---|---|---|---|
| 1 | $\|t_{11} \Rightarrow^{C,L} t_{21}\|$ | 0.568 | 0.575 |
| 2 | $\|t_{21} \Rightarrow^{C,L} t_{12}\|$ | 0.875 | 0.997 |
| 3 | $\|t_{12} \Rightarrow^{C,L} t_{24}\|$ | 0.581 | 0.989 |
| 4 | $\|t_{24} \Rightarrow^{C,L} t_{13}\|$ | 0.971 | 0.997 |
| 5 | $\|t_{13} \Rightarrow^{C,L} t_{25}\|$ | 0.808 | 0.837 |
| 6 | $\|t_{27} \Rightarrow^{C,L} t_{16}\|$ | 0.848 | 0.872 |
| 7 | $\|t_{12} \Rightarrow^{C,L} t_{25}\|$ | 0.095 | 0.050 |
| 8 | $\|t_{13} \Rightarrow^{C,L} t_{21}\|$ | 0.667 | 0.952 |

从图 2 可知，只有编码 7 代表的一组数据小于引入的阈值 $\theta = 0.500$，编码 7 代表的具有条件直接跟随关系的活动(变迁或特征)对分别是 $t_{12}$ 和 $t_{25}$。包含此活动对($t_{12}$ 和 $t_{25}$)的事件轨迹 $\sigma_7$，$\sigma_8$，$\sigma_9$ 被删除。而其他 7 组数据均大于引入的阈值 $\theta$，它们分别是编码 1，2，3，4，5，6 以及 8 所代表的数值。编码 1，2，3，4，5，6 和 8 代表的具有条件直接跟随关系的活动(变迁或特征)对分别是 $t_{11}$ 和 $t_{21}$、$t_{21}$ 和 $t_{12}$、$t_{12}$ 和 $t_{24}$、$t_{24}$ 和 $t_{13}$、$t_{13}$ 和 $t_{25}$、$t_{27}$ 和 $t_{16}$ 以及 $t_{13}$ 和 $t_{21}$。基于此，事件轨迹 $\sigma_6$ 中的交互特征对($t_{11}$ 和 $t_{21}$、$t_{21}$ 和 $t_{12}$、$t_{12}$ 和 $t_{24}$、$t_{24}$ 和 $t_{13}$、$t_{13}$ 和 $t_{27}$ 和 $t_{16}$)的条件依赖绝对值都大于阈值，$\sigma_6$ 被保存下来。在业务过程源模型 $M_O$ 的可执行轨迹中去掉 $\sigma_7$，$\sigma_8$，$\sigma_9$ 后，余下的事件轨迹集为优化事件轨迹集。

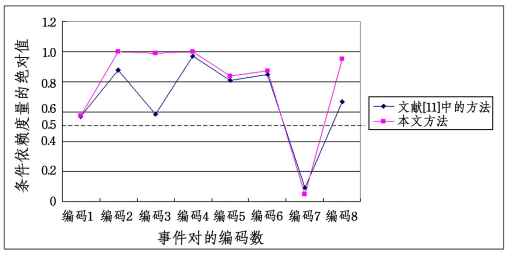


图 2　事件对的条件依赖度量值比较

Fig. 2　Comparison of conditional dependency measure of event pairs

利用文献[11]中的方法所得出的数据接近于阈值 $\theta$，数据点在阈值上下浮动的幅度小于本文方法的数据浮动幅度，不易于观察；而利用本文方法计算的数据与阈值的距离相差较大，可以更加直观形象地解决不频繁事件轨迹的去留问题，进而优化了事件日志。

依据算法 1 处理后的优化事件轨迹集，利用 $\alpha$ 算法挖掘业务过程优化通讯模型 $M_T$，模块网 $M'$ 与 $M''$ 之间的特征网 $M_f$ 即为模块网交互的信息通讯区域，$p_1$，$p_2$，$p_3$，$p_4$，$p_5$ 以及 $p_6$ 均是端口库所即两个模块网之间可以进行交互的信息流。其中，优化后的进行通讯的特征对之间的部分条件非频繁行为被移除，例如，特征(事件或变迁)对 $t_{11}$ 与 $t_{24}$ 之间的库所及流关系被过滤掉，从特征 $t_{25}$ 流向 $t_{13}$ 的信息流即 $t_{25}$ 与 $t_{13}$ 之间的库所及流关系被删除。然而，迹 $\sigma_6$ 记录的条件非频繁行为得以保留。具体的优化模型如图 3 所示。

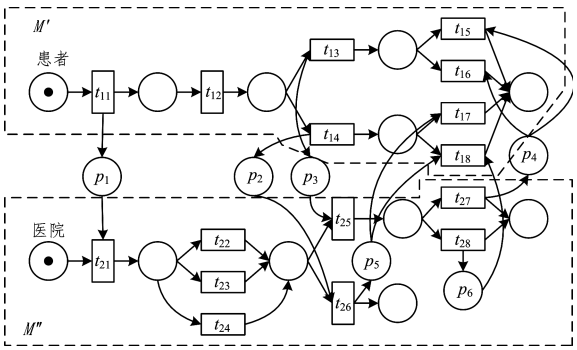


图 3　业务过程优化通讯模型 $M_T$

Fig. 3　Business process optimization communication model $M_T$

另外,已有的方法直接删除了条件非频繁行为,对频繁行为进行分析处理。而本文并没有全部去掉条件非频繁行为,而是适当地保留了部分含有重要信息的行为。相比于已有算法,本文方法在时间和空间效率上的代价会更大,基于此,本文没有在时间和空间上进行比较分析。

**结束语** 本文在已有研究的基础上,给出了基于通讯行为轮廓挖掘条件非频繁行为的优化方法。该方法以模块网之间的通讯行为轮廓为基础,首先给定业务过程源模型,查询其可执行事件日志,找出频数较低的事件轨迹并赋予其属性值。依据属性化的条件非频繁轨迹,计算不同模块网之间活动对的条件依赖度量值并进行比较,从而确定轨迹是否被保留或者删除,得到优化事件日志。依据优化事件轨迹集挖掘出业务流程优化通讯模型。

未来将以不同模块间的通讯行为轮廓理论为基础,对于事件数量较多的频数较低事件轨迹集,挖掘带有配置信息的非频繁行为,以提高业务过程系统的运作效率。

## 参 考 文 献


[1] VAN DER AALST W M P. Process mining data science in action(Second Edition)[M]. Springer,2016.

[2] WEERDT J D,BACKER M D,VANTHIENEN J,et al. A multi-dimensional quality assessment of state-of-the-art process discovery algorithms using real-life event logs[J]. Information Systems,2012,37(7):654-676.

[3] SURIADI S,ANDREWS R,TER HOFSTEDE A H M,et al. Event log imperfection patterns for process mining:Towards a systematic approach to cleaning event logs[J]. Information Systems,2017,64(C):132-150.

[4] 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.

[5] PONCE-DE-LEÓN H,CARMONA J,VANDEN BROUCKE S K L M. Incorporating negative information in process discovery[C]// International Conference on Business Process Management. Springer,2015:126-143.

[6] BELLODI E,RIGUZZI F,LAMMA E. Statistical relational learning for workflow mining[J]. Intelligent Data Analysis,2016,20(3):515-541.

[7] CONFORTI R,ROSA M L,HOFSTEDE T,et al. Filtering out infrequent behavior from business process event logs[J]. IEEE Transactions on Knowledge & Data Engineering,2017,29(2),300-314.

[8] ZELST S J V,DONGEN B F V,AALST W M P V D,et al. Discovering relaxed sound workflow nets using integer linear programming[J]. Computing,2017(4).

[9] LEEMANS S J J,FAHLAND D,VAN DER AALST W M P. Discovering block-structured process models from event logs containing infrequent behaviour[C]// International Conference on Business Process Management. Springer,Cham,2013:66-78.

[10] VAN DER WERF J M,KAATS E. Discovery of functional architectures from event logs[C]//PNSE@ Petri Nets. 2015:227-243.

[11] MANNHARDT F,DE LEONI M,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.

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


(上接第 309 页)


[8] LIU J X,MA T,CHEN J J. RELAX-Based Method for Aifcraft Fuel Consumption Performance Evaluation[J]. Electronics Optics & Control,2015,22(9):101-105. (in Chinese)
刘家学,马涛,陈静杰. 基于 RELAX 算法的飞机油耗性能估计方法[J]. 电光与控制,2015,22(9):101-105.

[9] CHEN J J,LI L Q. Fuel Consumption Classification of Aircraft in Descent Based on K-means Algorithm[J]. Measurement & Control Technology,2015,34(11):16-19. (in Chinese)
陈静杰,李吕琪. 基于 K-means 算法的飞机下降过程油耗分类[J]. 测控技术,2015,34(11):16-19.

[10] CHEN J J,XIAO G P. Analysis tool design of aircraft fuel consumption[J]. Computer Engineering and Design,2014,35(11):4012-4016. (in Chinese)
陈静杰,肖冠平. 飞机油耗分析工具设计[J]. 计算机工程与设计,2014,35(11):4012-4016.

[11] WANG X H,CHEN X F. A Support Vector Method for Modeling Civil Aircraft Fuel Consumption with ROC Optimization[C]// 2014 Second International Conference on Enterprise Systems. Shanghai,China,2014:112-116.

[12] AROCRA P,DEEPALI D,VARSHNEY S. Analysis of K-Means and K-Medodis Algorithm For Big Data[C]// International Conference on Information Security & Privacy(ICISP2015). Nagpur,INDIA,2015:507-512.

[13] HAN L S,XIANG L S,LIU X Y,et al. The K-medodis Algorithm with Initial Centers Optimized Based on a P System[J]. Journal of Information & Computational Science,2014,11(6):1765-1773.

[14] HE Y,PI D C. Iterative Imputation Algorithm Based on Reduced Relational Grade for Gene Expression Data[J]. Computer Science,2015,42(11):251-255,283. (in Chinese)
何云,皮德常. 基于精简关联度的基因表达数据迭代填补算法[J]. 计算机科学,2015,42(11):251-255,283.