

流程交换重复偏差的识别方法

王瑞莹, 杜玉越

(山东科技大学 信息科学与工程学院, 山东 青岛 266590)

摘要:针对现有流程偏差模式无法正确识别流程组合偏差的情况,提出一种流程组合偏差模式。该模式以构建偏差模式约束为前提,将活动替换、交换和重复偏差模式相互组合。以标准代价函数为基础,给出改进代价函数。根据组合偏差模式及改进代价函数,设计交换重复偏差模式。将交换重复偏差模式应用于签证申请流程,获得的校准结果可以识别流程中存在的偏差,验证了组合偏差模式能够有效识别流程组合偏差。

关键词:校准;重复偏差;组合偏差;过程挖掘;一致性检测;大数据

中图分类号:TP311.3

文献标志码:A

文章编号:1672-3767(2015)01-0035-07

An Identification Method of Process Swap and Repetition Deviation

Wang Ruiying, Du Yuyue

(College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao, Shandong 266590, China)

Abstract: A composite deviation pattern was proposed to improve the existing process deviation patterns which failed to properly identify the process deviations of high level. Action replacement, swap and repetition patterns were composed under a constructing constraint in the composite deviation pattern. A modified cost function was provided based on a standard cost function. According to the composed deviation pattern and modified cost function, a swap-repetition deviation pattern was designed and applied to a visa application handling process. In the alignment, the existing deviation of processes can be identified. This result illustrates that the composite deviation can effectively identify process deviation.

Key words: alignment; repetition deviation; composition deviation; process mining; conformance check; big data

大数据隐含着巨大的社会、经济及科研价值,已引起了各行各业的高度重视^[1]。大数据具有规模性(volume)、多样性(variety)和高速性(velocity)^[2]的特点,有效组织和利用大数据,将对社会经济和科学研究产生巨大推动作用。过程挖掘就是一种以分析得到的数据为基础,对其进行处理和组织的技术,主要包括过程发现、一致性检测和过程改进^[3]。一致性检测以日志和模型为输入,依据拟合度、精确度、泛化度和简洁度,度量过程模型的质量^[4-5]。仅对模型质量进行度量,不能帮助用户识别模型中的问题或对实际流程加强管理,所以,近年来对于一致性检测的研究,扩展到了对日志与模型的偏差部分做出诊断分析,以此为过程改进提供更多信息。

利用 Petri 网表示流程模型,以迹为单位,通过匹配迹中活动和网中变迁,可以表示迹与网之间的拟合程度,并以校准的形式记录匹配结果^[6]。在校准中可以清楚识别迹与网产生的偏差,同时,引入代价的概念,

收稿日期:2014-11-01

基金项目:国家自然科学基金项目(61170078, 61472228);教育部高等学校博士学科点专项科研基金博导类项目(20113718110004);青岛市科技计划基础研究项目(13-1-4-116-jch);山东科技大学科研创新团队支持计划项目(2011KYTD102)

作者简介:王瑞莹(1990—),女,山东聊城人,硕士研究生,主要从事流程挖掘、Petri 网等方面的研究。

杜玉越(1960—),男,山东聊城人,教授,博士生导师,主要从事 Petri 网、工作流、过程挖掘等方面的研究,本文通信作者。E-mail:yydu001@163.com

以最小代价为依据寻找最优校准,以此分析迹与网之间最可能的偏差,为了进一步分析偏差原因,采用高层次偏差模式的概念,以 Petri 网表示偏差模式,并在原网基础上增添偏差模式实例形成检测网,根据计算迹与检测网的最优校准识别引起偏差的最可能原因。

现有流程偏差模式通常分为三类,即活动替换、活动交换和活动重复偏差^[7-8]。但是,在现实流程中,还可能会产生三类情况组合下的偏差,而这些偏差也具有实际意义。已有偏差模式将三种偏差单独考虑,将导致无法正确识别组合偏差。因此,本研究将三类偏差模式进行组合,引入组合偏差模式,使其能够识别复杂的偏差情况。

1 基本概念

1.1 Petri 网

用来描述过程模型的形式化语言有很多种,比较常见的有 BPMN^[9] (business process modeling notation, 业务流程建模与标注), C-Nets^[10] (causal nets, 因果网) 以及 Petri 网^[11] 等。Petri 网是描述具有分布、并发、异步特征系统的有效工具,为 workflow 建模提供了形式化、图形化方法及分析技术。因此,本研究使用 Petri 网描述过程模型。

1.2 校准

校准是一组由迹中活动与网中变迁匹配得到的动作序列。匹配产生的动作分为四类:第一类为日志动作,表示在日志中发生的活动在模型中没有发生;第二类为模型动作,表示在模型中应该发生的活动在日志中没有被记录,即实际流程中没有发生;第三类为同步动作,表示在日志与模型中都发生了的活动;第四类为不合法动作,即除上述三类。

校准的表示如图 1 所示,第一行表示迹中活动序列,第二行表示网中完整引发序列。每一列是一个迹中活动与网中变迁通过相同标签进行匹配后产生的动作。图 1 中第 2 列表示模型动作,第 4 列表示日志动作,剩余列均表示同步动作。

$$\gamma_i = \begin{array}{|c|c|c|c|c|c|} \hline & register & >> & interview & check\ document & fingerprint & decision \\ \hline & register & fingerprint & interview & check\ document & >> & decision \\ \hline & t_1 & t_2 & t_3 & t_5 & & t_6 \\ \hline \end{array}$$

图 1 校准的表示

Fig. 1 The expression of alignments

标准代价函数是对校准中迹和模型的相似性度量,通过对校准中不同动作赋予不同代价(数值形式),表示迹和模型的相似程度。校准中同步活动代价设为数值 0,模型动作、日志动作代价设为 1,不合法动作代价设为 $+\infty$ ^[6]。

最优校准是基于代价函数计算迹与模型之间所有校准的代价后,代价最小的校准。

1.3 偏差模式

偏差模式是一个可以描述高层次偏差的 Petri 网片段,且偏差模式仅考虑该模式所代表的偏差。给定一条迹和一个以 Petri 网表示的业务流程,将偏差模式对应 Petri 网片段添加到原网上,以检测迹中是否出现该模式所对应的偏差。有三种基本偏差模式:活动替换偏差模式(用以检测活动之间相互替换的偏差模式),活动交换偏差模式(用以检测活动之间顺序发生交换的偏差模式),以及活动重复偏差模式(用以检测活动多次发生的偏差模式)。

在原网基础上增添偏差模式实例之后的网,称之为检测网。检测网仅用于识别表示实际流程的迹中是否出现增添的偏差模式所代表的偏差,且仅在进行校准时使用。

检测网代价函数是基于标准代价函数对校准中迹和检测网的相似性度量。迹与检测网之间的校准中,偏差模式中变迁对应的同步动作整体代价设为 1,模型动作代价设为无限大($+\infty$)。检测网中属于原网的

变迁所对应的动作产生的代价仍采用基本代价函数。

2 现有流程偏差模式存在的问题

流程偏差模式有活动替换、活动交换和活动重复偏差模式。对迹以及模型进行校准时,偏差模式只能识别它所描述的偏差。现实流程中可能会产生三类情况组合的偏差,而这些偏差也具有实际意义。由于已有偏差模式将三种偏差单独考虑,导致无法正确识别流程中的组合偏差。本研究采用申请签证简要流程进行说明,如图 2 所示。

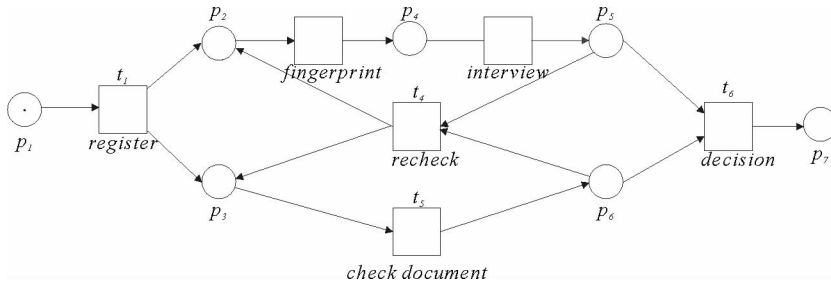


图 2 申请签证流程

Fig. 2 A process of visa application

申请签证流程以申请人递交申请材料并注册个人信息开始(*register*)。当一位工作人员负责审核申请人材料信息时(*check document*),另一位工作人员负责使用指纹机采集申请者指纹(*fingerprint*),并在此之后对申请人进行面谈(*interview*)。如果在申请过程中发现任何可疑信息,一名管理人员可能会对申请者进行复核(*recheck*),复核可能包括审核申请者的所有材料信息。流程以管理人员对申请作出决定为结束(*decision*)。

给定迹 $\sigma_1 = \langle register, interview, fingerprint, fingerprint, fingerprint, fingerprint, recheck, decision \rangle$ 以及图 2 所示过程网。假设当前认为可能发生的高层次偏差包含:*fingerprint* 与 *interview* 交换顺序;*recheck* 替换 *check document*; *fingerprint* 重复执行。

图 3 为在原网基础上增添上述三种偏差模式实例后的检测网,其中, t_{11} 到 t_{14} 部分的 Petri 网片段表示重复偏差模式, t_{16} 和 t_{17} 部分的 Petri 网片段表示交换偏差模式, t_{15} 部分的 Petri 网片段表示替换偏差模式。虚线框内给出了偏差模式中变迁赋予的代价。

图 4 描述了迹与检测网基于代价函数的最优校准。通过查找偏差模式实例中初始变迁对应的同步活动,可以从最优校准中识别出所有高层次偏差。

图 4 中存在两列相对应的初始变迁。第 3 列的变迁 t_{11} 为重复偏差模式的初始变迁,第 9 列的变迁 t_{15} 为替换模式的初始变迁。通过分析,图 4 所示校准中包含两个高层次偏差:*fingerprint* 的重复执行;*recheck* 替换 *check document*。另外,校准中同时表明了两个低层次偏差:变迁 *interview* 对应的日志动作;变迁 *interview* 对应的模型动作。该最优校准相应代价为 4。

在该校准中,识别出该条迹与模型不匹配的原因是:①活动 *fingerprint* 发生了重复;②*recheck document* 替换了 *check document*;③*interview* 在模型中不该发生,但发生于日志中;④*interview* 在模型中应该发生,但日志中却没有记录,由此产生的校准代价为 4。但通过分析迹 $\sigma_2 = \langle register, interview, fingerprint, fingerprint, fingerprint, fingerprint, recheck, decision \rangle$,可以对偏差原因做出以下推测:①*interview* 与 *fingerprint* 发生了顺序交换;②*fingerprint* 重复发生多次;③*recheck* 替换 *check document* 发生。根据检测网代价函数,由此产生的校准代价为 3。

因此,迹与模型产生偏差应该这是由于 *interview* 与 *fingerprint* 先发生了顺序交换,其后 *fingerprint* 重复发生多次,并且 *recheck* 替换了 *check document*。但是,由于重复模式及替换模式相对独立,导致其在检测

网中既定的顺序不能检测出先交换再重复的偏差模式,即交换重复组合模式。

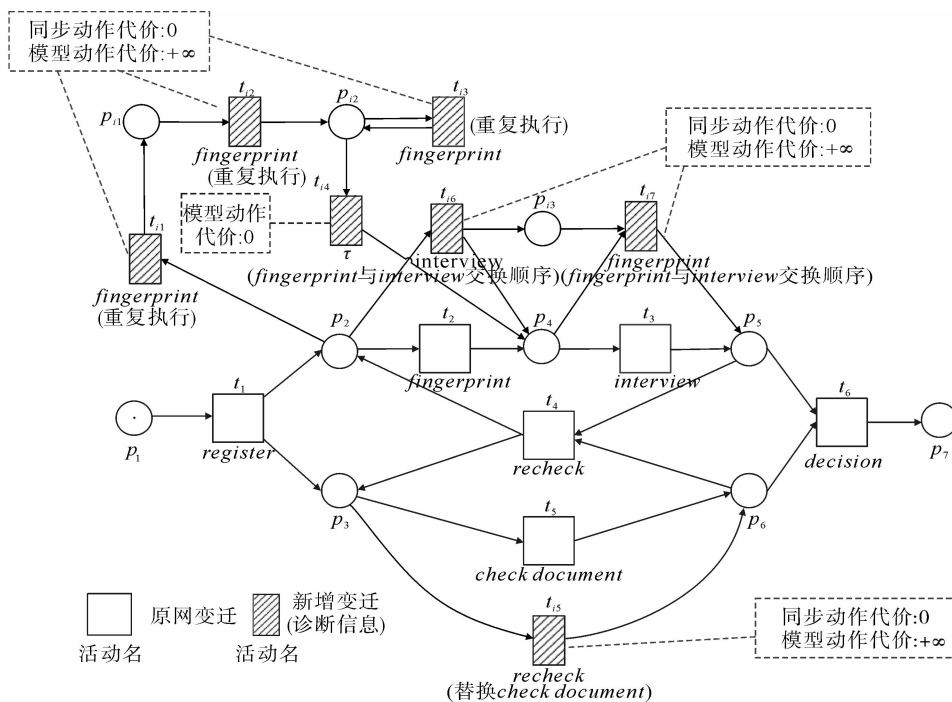


图 3 加入三种偏差模式实例后的检测网

Fig. 3 The detecting net with pattern instances of three deviation patterns

迹	reg	int	fp	fp	fp	fp	>>	>>	reck	dec
模型	reg	>>	fp	fp	fp	fp	t ₁₄	t ₃	reck	dec
	t ₁		t ₁₁	t ₁₂	t ₁₃	t ₁₃			t ₁₅	t ₆
诊断	—	—	fingerprint 重复执行				—	check document 被 recheck 替换		

注:reg—register;fp—fingerprint;int—interview;dec—decision;reck—recheck

图 4 迹 σ_1 与图 3 网之间的最优校准

Fig. 4 An optimal alignment between σ_1 and the detecting net in Fig. 3

3 组合偏差模式

3.1 相关概念

给定一条迹、一个以 Petri 网描述的流程模型和一个偏差模式,任何由给定偏差模式实例所描述的偏差,将从迹与网之间的校准转换为高层次诊断。为此,引入偏差模式约束:偏差模式应仅包含唯一的初始变迁和终止变迁。任何单进单出偏差模式的引发序列,起始于初始变迁,结束于终止变迁,并且中间不再包含初始/终止变迁。在加入偏差模式后,原网的可达标识应当全部保留。因此,引入偏差模式约束后,通过校准中初始变迁的同步活动和终止变迁的同步活动,任何偏差都能够被标记和识别。

在满足构建偏差模式约束的前提下,将现有的三种基本偏差模式组合所形成的偏差模式称为流程组合偏差模式。

改进代价函数是引入由两种基本偏差模式构建的组合偏差模式后,基于标准代价函数和改进代价函数对校准中迹和检测网的相似性度量。迹与检测网之间的校准中,组合偏差模式中变迁所对应的同步动作代

价设为 2,模型动作代价设为无限大($+\infty$)。检测网中属于原网的变迁所对应动作产生的代价,仍采用基本代价函数。通过这种方式增添组合偏差模式实例后的检测网与迹产生的最优校准中,不会仅包含组合偏差模式实例中的部分变迁,即包括全部组合偏差模式实例中的变迁或全部不包括。

3.2 交换重复偏差模式

交换重复偏差模式是一种流程组合偏差模式,用于检测实际流程是否出现了由于活动交换且重复所导致的偏差。

对于突发状况,一个活动在与其它活动交换顺序后可能还需要执行多次。例如,在申请签证流程中,负责采集指纹及面试阶段的工作人员首先对申请人进行面试,之后再指纹录入,并且在指纹录入阶段由于指纹识别机的故障可能造成多次采集指纹。例如,迹 $\sigma_2 = \langle register, interview, fingerprint, fingerprint, fingerprint, check\ document, decision \rangle$ 是这种情况下产生的一条迹。可以清楚地看出活动 *fingerprint* 与 *interview* 发生了顺序交换,并且在此之后 *fingerprint* 重复发生了多次。基于交换重复偏差模式,可以检测实际流程是否出现活动先交换又重复发生的情况。

首先在原网增加交换重复偏差模式实例构成检测网,然后计算迹与检测网之间的最优校准。

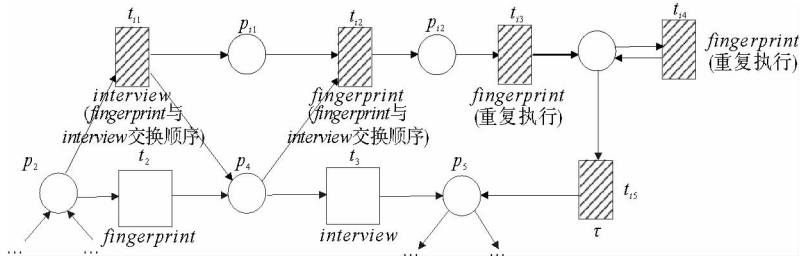


图 5 交换重复模式的检测网

Fig. 5 The detecting net with swap-repetition patterns

图 5 为加入交换重复模式的检测网。变迁 t_{11} 表明交换重复模式实例的开始,变迁 t_{11} 与变迁 t_{12} 描述原网中变迁 *fingerprint*, *interview* 之间的顺序交换。变迁 t_{13} 紧跟变迁 t_{12} 之后发生,表明 *fingerprint* 的重复发生,即保证变迁 *fingerprint* 至少发生两次。变迁 t_{14} 描述变迁 *fingerprint* 可以重复发生不定次数。变迁 t_{15} 标记了此模式的结束。

图 6 给出迹 σ_2 与检测网之间根据改进代价函数产生的最优校准。其中,第 2 列中变迁 t_{11} 为组合偏差模式的初始变迁,第 7 列中变迁 t_{15} 为组合偏差模式的终止变迁。在校准中变迁 t_{11} 与 t_{15} 对应的同步动作,说明迹 σ_2 确实存在该组合模式所代表的偏差,即先交换再重复的偏差形式。根据改进代价函数,校准对应代价为 2。

迹	<i>reg</i>	<i>int</i>	<i>fp</i>	<i>fp</i>	<i>fp</i>	<i>fp</i>	>>	<i>ck</i>	<i>dec</i>
模型	t_1	t_{11}	t_{12}	t_{13}	t_{14}	t_{14}	t_{15}	t_5	t_6
诊断	—	fingerprint与interview交换		fingerprint重复执行			—	—	—
组合偏差									

注:reg-register;fp-fingerprint;int-interview;dec-decision;ck-check document;reck-recheck

图 6 迹 σ_2 与图 5 中检测网之间的最优校准

Fig. 6 An optimal alignment between σ_2 and detecting net in Fig. 5

对图 5 中的交换且重复偏差的实例,在交换重复组合偏差模式中抽象为图 7。

图 7 中 tp_1 和 tp_2 表示 X 与 Y 交换顺序, tp_3 和 tp_4 表示 X 的循环部分, tp_5 作为交换重复偏差模式的终止变迁。

针对第 2 节中的问题,使用交换重复组合偏差模式解决。首先,为流程模型加入设计的交换重复组合偏差模式实例,形成检测网。其次,计算检测网与迹之间的最优校准。

在图 3 中增添交换重复组合偏差模式后形成的检测网,如图 8 所示。图 8 省略了重复偏差模式和交换偏差模式的单独表示。基于改进代价函数,对迹 σ_1 与检测网进行匹配,得到最优校准如图 9 所示。

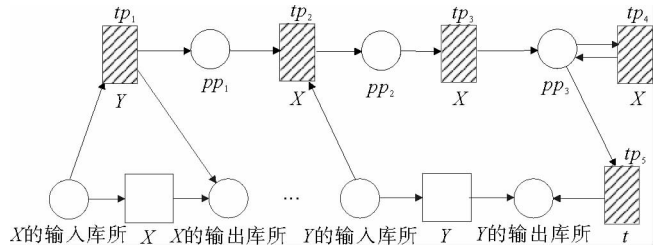


图 7 交换重复偏差模式

Fig. 7 Swap-repetition deviation pattern

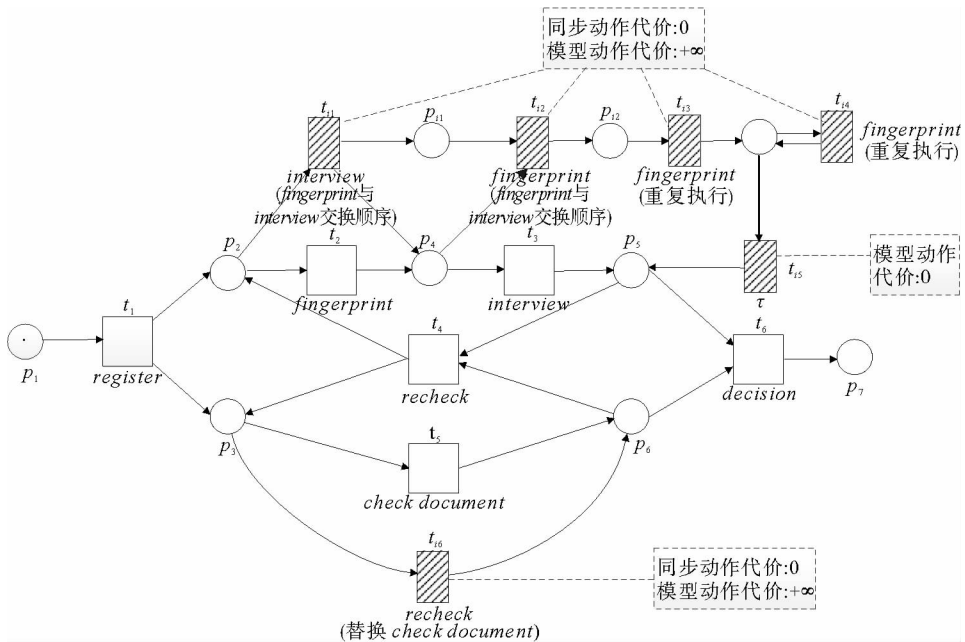


图 8 组合偏差模式的检测网

Fig. 8 The detecting net with composite deviation patterns

迹	reg	int	fp	fp	fp	fp	>>	reck	dec
模型	reg	int	fp	fp	fp	fp	t ₁₅	reck	dec
	t ₁	t ₁₁	t ₁₂	t ₁₃	t ₁₄	t ₁₄	t ₁₅	t ₁₅	t ₆
诊断	-	fingerprint与interview交换		fingerprint重复执行			check document被reck替换	-	-
		组合偏差							

注:reg-register;fp-fingerprint;int-interview;dec-decision;ck-check document;reck-recheck

图 9 迹 σ_1 与图 8 中检测网的最优校准

Fig. 9 The alignment between σ_1 and the detecting net in Fig. 9

图 9 识别出三种偏差:①interview 与 fingerprint 发生了顺序交换,②fingerprint 重复发生 3 次,③re-check 替换 check document。

根据改进代价函数,图9中校准的代价为3,较之于图4校准的代价为4更小。所以,图9中校准同时也是最优校准。根据图9的最优校准,可以分析出导致迹 σ_1 与模型发生偏差的根本原因是 *interview* 与 *fingerprint* 先发生了顺序交换,然后 *fingerprint* 重复发生多次,并且 *recheck* 替换 *check document*。在仅引入替换、交换和重复偏差模式的检测网与迹得到的校准中(图4),单个偏差模式对这种原因是不能识别的。因此,使用组合偏差模式,可以查找出更根本或实际的偏差原因,从而证明了组合偏差模式的实用性和有效性。

4 结束语

分析已有流程偏差模式不能识别流程中组合偏差的情况,提出一种流程组合偏差模式。为确保组合偏差模式实例所描述的偏差能够将迹与网之间的校准转换为诊断信息,引入一种偏差模式约束,并以此约束为前提,将活动替换、活动交换和活动重复偏差模式组合,形成组合偏差模式。以标准代价函数为基础,给出一种改进代价函数。结合组合偏差模式和改进代价函数,设计一种交换重复偏差模式。最后,将设计的交换重复偏差模式应用于实例分析中,解决了基本偏差模式无法识别组合偏差的问题,例证了组合偏差模式的实用性和有效性。下一步工作,将对流程组合偏差模式进行形式化描述,并结合其他流程组合偏差情况,丰富流程组合偏差模式的类型。

参考文献:

- [1]Labrinidis A, Jagadish H V. Challenges and opportunities with big data[J]. Proceedings of the VLDB Endowment, 2012, 5(12): 2032-2033.
- [2]Marx V. Biology: The big challenges of big data[J]. Nature, 2013, 498(7453): 255-260.
- [3]van der Aalst W, Weijters A. Process mining: A research agenda[J]. Computers in Industry, 2004, 53(3): 231-244.
- [4]王嘉. 面向过程挖掘的一致性分析方法研究[D]. 长沙: 国防科学技术大学, 2009: 25-26.
- [5]马璐璐. 流程发掘算法综合评价框架的设计与实现[D]. 北京: 北京交通大学, 2013: 32-34.
- [6]Adriansyah A. Aligning observed and modeled behavior[D]. Eindhoven: Eindhoven University of Technology, 2014: 99-110.
- [7]Depaire B, Swinnen J, Jans M, et al. A process deviation analysis framework[C]//Business Process Management Workshops. Berlin Heidelberg: Springer, 2013: 701-706.
- [8]Swinnen J, Depaire B, Jans M J, et al. A process deviation analysis: A case study[C]//Business Process Management Workshops. Berlin Heidelberg: Springer, 2012: 87-98.
- [9]Aguilar-Saven R S. Business process modelling: Review and framework[J]. International Journal of Production Economics, 2004, 90(2): 129-149.
- [10]van der Aalst W, Adriansyah A, van Dongen B. Causal nets: A modeling language tailored towards process discovery[M]//CONCUR 2011-Concurrency Theory. Berlin Heidelberg: Springer, 2011: 28-42.
- [11]Murata T. Petri nets: Properties, analysis and applications[J]. Proceedings of the IEEE, 1989, 77(4): 541-580.

(责任编辑: 吕文红)