基于偏差约减的大数据交易模型分析与修复方法

郭 艺[1],叶 剑[2,3],张 鹏[1]
(1. 山东科技大学,山东青岛 266590;2. 中国科学院计算技术研究所,北京 100190;
3. 移动计算与新型终端北京市重点实验室,北京 100190)

摘 要: 大数据交易是促进数据流通和提升数据价值的关键环节. 实现大数据交易的过程优化对于构建高效和鲁棒的交易平台至关重要. 大数据交易是典型的复杂过程模型,传统的模型修复方法无法有效发现和约减流程执行与流程规则之间存在的偏差. 本文提出了一种基于偏差约减的大数据交易模型修复方法,通过过程模型的可达标识图发现事件日志与模型之间的偏差关系,对事件日志与模型之间偏差进行约减,实现基于有效偏差的模型修复. 该方法应用于天元大数据网大数据平台,通过与基于模型校准和基于迭代的修复方法进行对比实验,对修复结果开展模型拟合度、精确度、简洁度及时间复杂度评估,验证了方法的有效性.

关键词: 大数据交易; 模型修复; 模型评估; 偏差约减
中图分类号: TP311 **文献标识码:** A **文章编号:** 0372-2112 (2018)07-1754-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.07.031

Analysis and Repair of Big Data Transaction Model Based on Deviation Reduction

GUO Yi[1], YE Jian[2,3], ZHANG Peng[1]
(1. *Shandong University of Science and Technology, Qingdao, Shandong* 266590, *China*;
2. *Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190, *China*;
3. *The Beijing Key Laboratory of Mobile Computing and Pervasive Device, Beijing* 100190, *China*)

Abstract: Big data transaction is a key point of promoting data circulation and data value. It is important for building efficient and robust trading platform to optimize process of big data transaction. Big data transaction is a typical complex process model, which makes the traditional model repair method not able to effectively discover and reduce the deviation between process execution and process rules. This paper proposes an approach of repairing big data transaction model based on deviation reduction. With the help of the reachable marking graph, the approach discovers the deviation between the event log and the process model found, reduces the deviation between the event log and the model, and gets the model repaired based on the effective deviation. At the end of this paper, the proposed approach is used in the Tianyuan big data platform to verify the effectiveness. In comparison experiments of those repair methods based on model alignment and the iteration, the effect of repairing is evaluated from the aspects of fitness, precision, simplicity and time complexity. The evaluation shows that the proposed approach has an advantage over existing methods.

Key words: big data transaction; model repair; model evaluation; deviation reduction

1 引言

随着海量数据的出现,数据的价值得到越来越多行业和领域的认可,大数据交易成为一种新型的数据流通手段,大数据交易平台的构建过程中产生了自有数据出售,数据加工转售以及数据交易平台服务等多种交易模式[1~3]. 如何及时发现并解决大数据交易流程中实际执行与业务规则之间的偏差是大数据交易过程面临的巨大挑战. 业务流程管理[4~7]通过对业务流程的整个生命周期进行建模、管理、监控和优化,来实现对各

收稿日期:2017-08-24;修回日期:2017-12-24;责任编辑:蓝红杰
基金项目:国家重点研发计划(No. 2016YFB1001105);国家自然科学基金(No. 61401040);工信部 2016 年集成制造系统集成项目和移动计算与新型终端北京市重点实验室研究基金

种业务环节的整合的全面管理. 作为业务流程管理的重要环节,过程挖掘[8~11]能够发现大数据交易业务流程存在的问题和瓶颈,以便对交易过程模型不断地优化. 模型修复是过程挖掘的重要部分,通过修改模型使新的模型更好的反映现实业务.

现有的模型修复方法主要是基于模型校准[12]来发现日志与模型之间的偏差以达到模型修复的目的,通过分解日志的方式[13,15]发现偏差,基于模型修复建议[14]达到修复模型的目的. 迭代的过程发现技术解决的问题类似于模型修复问题,通过对比日志与模型中的关系[22,23],基于关联规则[24]针对关系的不同点执行替换操作实现模型修复. 基于模型校准的模型修复方法和迭代的过程发现技术对模型修复具有一定的指导和借鉴意义,但是大数据交易模型修复过程中,在日志与模型拟合的情况下,当交易模型中多个并行分支之间存在偏差时,基于模型校准的模型修复方法无法准确的发现并解决偏差,基于迭代的过程发现技术执行过程复杂并且无法保证修复之后日志与模型之间的一致性.

针对上述问题,本文以大数据交易流程为切入点,在基于模型校准的模型修复方法和迭代过程发现技术的基础上,提出了基于偏差约减的模型修复方法. 该方法借助可达标识图的概念,从控制流的角度发现并分析流程执行过程中出现的偏差. 本文采用一致性分析标准[19~21],从拟合度、精确度和简洁度三个方面对方法的修复效果进行验证. 通过将该方法用于天元数据网的大数据交易平台的流程形式化分析,充分验证了方法的有效性.

2　相关概念

2.1　多重集、序列

$\mathcal{A}$ 是一个集合,存在映射 $\mathcal{B}$,$\mathcal{B}:\mathcal{A}\to\mathcal{B}(\mathcal{A})$,其中 $\mathcal{B}(\mathcal{A})$ 代表集合 $\mathcal{A}$ 上所有多重集的集合. 例如 $\mathcal{A}=\{a,b,c\}$,多重集 $B_i\in\mathcal{B}(\mathcal{A})$,$B_1=[\,]$,$B_2=[a]$,$B_3=[a^2,c,b,b^2]$. 序列 $\sigma=\langle a_1,a_2,\cdots,a_n\rangle\in\mathcal{A}^*$,$a_i\in\mathcal{A}$,$n\in\mathbb{N}_0$,$|\sigma|$ 表示序列 σ 的长度,$|\sigma|=n$.

2.2　标识 Petri 网

四元组 $N=(P,T,F,M)$ 称为一个标识 Petri 网,P 称为库所集,T 称为变迁集,F 是网 N 的流关系. 映射 $M:P\to\{0,1,2,\cdots\}$ 称为网 N 的一个标识,一个网系统有一个初始标识,记为 M_0. 标识 Petri 网具有下面变迁发生规则:

(1)变迁 $t\in T$,如果 $\forall p\in P:p\in\cdot t\to M(p)\geqslant 1$ 则说变迁 t 在标识 M 有发生权,记为 $M[t>$.

(2)若 $M[t>$,则在标识 M 下,变迁 t 可以发生,从标识 M 发生变迁 t 得到一个新的标识 M'(记为 $M[t>M'$),对 $\forall p\in P$,

$$M'(p)=\begin{cases}M(p)-1, & \text{若 } p\in\cdot t-t\cdot\\ M(p)+1, & \text{若 } p\in t\cdot-\cdot t\\ M(p), & \text{其他}\end{cases}$$

2.3　有界 Petri 网

设 $N=(P,T,F,M)$ 为一个 Petri 网,$p\in P$. 从 M 可达的一切标识的集合记为 $R(M)$. 若存在正整数 B,使得 $\forall M\in R(M_0):M(p)\leqslant B$,则称为库所 p 是有界的. 如果对于每个 $p\in P$ 都是有界的,则称为网 N 为有界 Petri 网.

2.4　可达标识图

设 $N=(P,T,F,M_0)$ 为一个有界 Petri 网. N 的可达标识图定义为三元组 $RG(N)=(R(M_0),e,p)$,其中

$e=\{(M_i,M_j)\mid M_i,M_j\in R(M_0),\exists t_k\in T:M_i[t_k>M_j\}$

$p:P\to T,p(M_i,M_j)=t_k\Leftrightarrow M_i[t_k>M_j$

2.5　事件日志

事件日志由案例组成,案例由事件组成. 案例中的事件用轨迹的形式来表示,即(唯一的)事件的序列. 设集合 $\mathcal{A}$ 代表事件日志中所有标签的集合,轨迹中的每一个标签 $\nu\in\mathcal{A}^*$ 代表一个事件,如事件日志 $L=[<a,b,d>^3,<a,c,d>^{10}]$,数字代表日志中对应案例的个数,$\lambda$ 代表标签与变迁之间的映射,如果网 N 中的变迁在 $\mathcal{A}$ 中没有对应的标签,则记为 $\lambda(t)=\tau$,存在 $t\in T,\lambda:T\to A\cup\{\tau\}$,如果 $\lambda(t)\neq\tau$,则 $\lambda(t)\in A$.

3　基于可达标识图的偏差发现与分析

实际的业务流程往往随着时间的变化不断地发生变化,为了适应不断变化的情况,通过模型修复算法修改模型使其更好的反映现实业务. 在大数据交易流程中,由于多个并行分支的存在,当日志与模型存在偏差时,已有的模型修复方法无法发现偏差. 发现日志与模型之间的偏差是模型修复的前提,本节基于可达标识图发现日志与模型之间的偏差.

发现偏差过程以大数据交易流程中数据加工流程为例,采用日志 $L=\{<a,b,d,c,e,f>^{11},<a,b,c,d,e,f>^{24},<a,c,b,d,e,f>^{11},<a,b,c,e,d,f>^{2}\}$ 其中事件含义如下:a 进入数据中心,b 数据选择,c 工具选择,d 数据审批,e 工具审批,f 数据加工. 根据文献[16]中 α 算法在日志 L 的基础上执行模型发现算法最终得到数据加工流程如图 1 网 N_1 所示. 目前应用最为广泛的偏差发现方法是通过模型校准的方式将日志中所有的轨迹在模型中重演,在重演的过程中发现日志与模型中的偏差,对于不同类型的偏差采用不同的方法执行模型修复.

基于模型校准的方式将事件日志 L 在网 N_1 中重演得到如表 1 所示 4 个模型校准的结果.

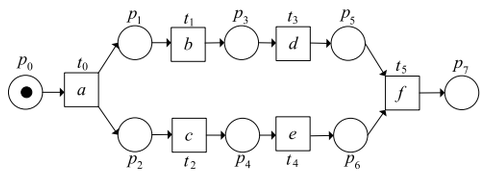

图1 Petri网N_1

表1 模型校准

γ_1:

a	b	d	c	e	f
a	b	d	c	e	f
t_0	t_1	t_3	t_2	t_4	t_5

γ_2:

a	b	c	d	e	f
a	b	c	d	e	f
t_0	t_1	t_2	t_3	t_4	t_5

γ_3:

a	c	b	d	e	f
a	c	b	d	e	f
t_0	t_2	t_1	t_3	t_4	t_5

γ_4:

a	b	c	e	d	f
a	b	c	e	d	f
t_0	t_1	t_2	t_4	t_3	t_5

从模型校准的结果可以看出模型校准过程中日志与模型之间没有发现偏差,但是通过分析日志可以发现除了在极少数事件日志中,事件 e 与事件 b 以及事件 e 与事件 d 是非并行关系,但是观察模型发现事件 e 与事件 b、事件 d 在过程模型中分别在两个并行分支中,所以模型校准无法检测出日志与模型之间的偏差.

为了避免上述情况,从模型可达标识图的角度进行分析发现日志与模型之间的偏差,假设网 N_1 的初始标识 $M_0=[1,0,0,0,0,0,0,0]$,基于可达标识图的概念,进一步给出了邻边关系集,子轨迹,最小状态偏差率以及偏差集的定义.

定义1(邻边关系集 R_c) 设网 $N=(P,T,F,M_0)$ 为有界 Petri 网,根据网 N 的可达标识图 $RG(N)=(R(M_0),E,P)$定义邻边关系集 $R_c=\{<t_a,M_i,t_b>\in T\times R(M_0)\times T\mid \exists t_a,t_b\in T,\exists M_i,M_j,M_k\in R(M_0):M_j[t_a>M_i[t_b>M_k\}$,令 $r=<t_a,M_i,t_b>$ 为邻边关系集 R_c 中的元素,定义关系前键 $\cdot r=t_a$,关系后键 $r\cdot=t_b$.

定义2(子轨迹) 日志 L 为事件轨迹的集合,设 σ 为事件日志 L 中的一条轨迹,$\sigma=<e_1,e_2,\cdots,e_n>\in L,n=len(\sigma)>1$,定义 $\sigma^{[i,j]}=<e_i,..,e_j>$ 其中 $i,j\in[1..n]$ 且 $i<j$,表示整条轨迹 σ 中从第 i 个事件到第 j 个事件之间所对应的子轨迹.

定义3(最小状态偏差率 P_{msd}) 设网 $N=(P,T,F,M_0)$ 为有界 Petri 网,日志 L 为事件轨迹的集合,令 count(M)为日志遍历过程中状态 M 出现的次数,$M\in R(M_0)$,max(M_L)表示日志遍历过程中单个状态出现的最大次数,状态偏差率 $P_M=\text{count}(M)/\max(M_L)$,其中 P_{msd}为日志遍历过程中允许的最小状态偏差率. 以日志 L 为例,在日志 L 的遍历过程中,状态[1,0,0,0,0,0,0,0]出现48次,同时状态[0,1,1,0,0,0,0,0],[0,0,0,0,0,1,1,0]与状态[0,0,0,0,0,0,0,1]也分别出现48次,在日志的遍历过程中统计得单个状态出现的最大次数为,即 $\max(M_L)=48$,所以状态[1,0,0,0,0,0,0,0]的状态偏差率 $P_{[1,0,0,0,0,0,0,0]}=\frac{48}{48}=1$ 同样从遍历过程中可得状态[0,0,1,1,0,0,0,0],[01,0,0,0,0,1,0]与状态[0,0,0,1,0,0,1,0]的状态偏差率分别 $P_{[0,0,1,1,0,0,0,0]}=\frac{37}{48}\approx 0.7708$, $P_{[0,1,0,0,0,0,1,0]}=\frac{0}{48}=0$, $P_{[0,0,0,1,0,0,1,0]}=\frac{2}{48}\approx 0.0417$.

定义4(偏差集 R_d) 设网 $N=(P,T,F,M_0)$ 为有界 Petri 网,$RG(N)=(R(M_0),E,P)$为网 N 对应的可达标识图,R_c 为网 N 对应的邻边关系集,定义偏差集 $R_d=\{<t_a,M_i,t_b>\in T\times R(M_0)\times T\mid \exists t_a,t_b\in T,\exists M_i,M_j,M_k\in R(M_0):M_j[t_a>M_i[t_b>M_k,P_{M_i}=\frac{\text{count}(M_i)}{\max(M_L)}<P_{msd}\}$.

偏差集 R_d 是通过事件日志在模型对应的可达标识图中迭代得到的. 根据子轨迹的定义将事件日志中的轨迹划分成若干子轨迹的形式,依次将子轨迹在可达标识图中遍历,在遍历过程中以最小状态偏差率作为约束条件,最终得到邻边关系集 R_c 的子集 R_d. 算法1是偏差集 R_d 的发现算法.

算法1 偏差集 R_d 发现算法

Input 邻边关系集 R_c,事件日志 L,可达标识图 $RG(N)$
Output 偏差集 R_d
$M_a=M_0,M_b=\mathbf{0},n=\text{count}(M_L)$
for each σ in L do
  for each $<e_i,e_j>$ in σ from $i=0,j=1$ to $i=len(\sigma)-1,j=len(\sigma)$ do
    $M_a[\lambda(e_i)>M_b$
    if$(\lambda(e_i),M_b,\lambda(e_j))$ in R_c then
      $R_d\leftarrow<\lambda(e_i),M_b,\lambda(e_j)>$
    $M_a=M_b,M_b=\mathbf{0}$,
    $n++$
  for each r in R_d do
    $m=\text{count}(M)$ in r
    if $(\frac{m}{\max(M_L)})>P_{msd}$
    then remove r in R_d
  return R_d

定义5(邻边关系差集 R_p) 设网 $N=(P,T,F,M_0)$ 为有界 Petri 网,$RG(N)=(R(M_0),E,P)$为网 N 的可达标识图,R_c 为网 N 对应的邻边关系集,R_d 为日志 L 对应的偏差集,定义邻边关系差集

$R_p=\{<t_a,M_i,t_b>\in T\times R(M_0)\times T\mid \exists t_a,t_b\in T,\exists M_i,M_j,M_k\in R(M_0):M_j[t_a>M_i[t_b>M_k,R_p=R_c-R_d\}$.

可以看出邻边关系差集 R_p 是邻边关系集 R_c 与偏差集 R_d 的差集,在计算偏差集的过程中假设最小状态偏差率为 0.05,由以上定义和方法发现只有状态[0,1,0,0,0,0,1,0]与状态[0,0,0,1,0,0,1,0]符合偏差集的定义,因为 $P_{[0,1,0,0,0,0,1,0]} = 0/48 = 0 < 0.05$, $P_{[0,0,0,1,0,0,1,0]} = 2/48 \approx 0.0417 < 0.05$,由网 N_1 的状态标识图以及偏差集可得如图 2 所示的偏差标记图,图中虚线表示的即为偏差状态以及与偏差状态所关联的边,所以网 N_1 对应的偏差集 $R_d = \{ < t_4, [0,1,0,0,0,0,1,0], t_1 >, < t_4, [0,0,0,1,0,0,1,0], t_3 >, < t_1, [0,0,0,1,0,0,1,0], t_3 > \}$. 偏差标记图如图 2 所示,虚线表示对应的偏差集 R_d.

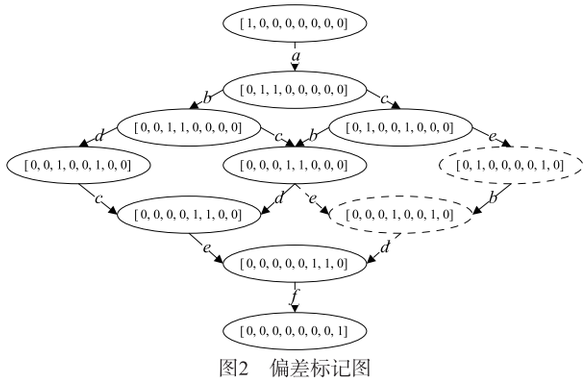

图2　偏差标记图

4　基于偏差约减的模型修复

在模型修复的过程中,基于可达标识图的偏差发现方法解决了无法发现日志与模型之间偏差的问题,同时得到日志与模型之间的偏差集作为模型修复的前提. 本节利用偏差集,提出基于偏差约减的模型修复方法,执行具体的模型修复操作.

定义 6(事件轨迹映射关系 π)　日志 L 为事件轨迹的集合,设 σ 为事件日志 L 中的一条轨迹,定义事件轨迹之间的映射关系 π,设 e_i 为轨迹 σ 中第 i 个事件,则 $e_i = \pi_i(\sigma)$.

定义 7(邻边关系差序列集 *PS*)　设 $N = (P, T, F, M_0)$ 为有界 Petri 网,$RG(N) = (R(M_0), E, P)$ 为网 N 对应的可达标识图,R_p 为网 N 对应的邻边关系差集,定义模型邻边关系差序列集 $PS = \{ < t_a, t_b > | \exists t_a, t_b \in T, \exists M_i \in R(M_0), \exists r \in R_p : r \cdot = t_a, \cdot r = t_b \}$

定义 8(修复序列集 *RS*)　设 $N = (P, T, F, M_0)$ 为有界 Petri 网,$RG(N) = (R(M_0), E, P)$ 为网 N 对应的可达标识图,R_d 为网 N 对应的偏差集,定义模型修复序列集 $RS = \{ < t_a, t_b > | \exists t_a, t_b \in T, \exists M_i \in R(M_0), \exists r \in R_d : r \cdot = t_a, \cdot r = t_b, \nexists r' \in R_p : r \cdot = r' \cdot, \cdot r = \cdot r' \}$.

为了执行模型修复操作,基于邻边关系差集 R_p 与偏差集 R_d 分别定义了邻边关系差序列集 PS 与修复序列集 RS 用于模型修复操作,算法 2 是基于偏差约减的模型修复算法.

算法 2　基于偏差约减的模型修复算法

Input 有界 Petri 网 $N = (P, T, F, M_0)$, 偏差集 R_d, 邻边关系差序列集 PS
Output $N_{repaired}(P', T, F', M_0)$
$RS' = \varnothing$
for each r in R_d do
  if $< r \cdot, \cdot r >$ not in PS then
    $RS = RS \cup \{ < r \cdot, \cdot r > \}$ //遍历 R_d 得到修复序列集 RS
for each σ_i in RS from $i = 1$ to $i = (n-1)$ do
    if σ_i in PS then continue
  for each σ_j in RS from $j = i+1$ to $j = n$ do
    if σ_j in PS then continue
    else if $\pi_2(\sigma_i) = = \pi_2(\sigma_j)$ then
//如果两个序列中第二个元素相同但是第一个元素存在次序关系,只保留一个序列
      if $< \pi_1(\sigma_i), \pi_1(\sigma_j) >$ in PS and $< \pi_1(\sigma_j), \pi_1(\sigma_i) >$ not in PS
      then $RS' = RS' \cup \{\sigma_j\} - \{\sigma_i\}$
      else if $< \pi_1(\sigma_j), \pi_1(\sigma_i) >$ in PS and $< \pi_1(\sigma_i), \pi_1(\sigma_j) >$ not in PS
      then $RS' = RS' \cup \{\sigma_i\} - \{\sigma_j\}$
//如果两个序列第二个元素相同但是第一个元素不存在次序关系,保留两个序列
      else $RS' = RS' \cup \{\sigma_i\} \cup \{\sigma_j\}$
//如果两个序列第二个元素不相同,保留两个序列
      else $RS' = RS' \cup \{\sigma_i\} \cup \{\sigma_j\}$
for each σ in RS' do//遍历 RS' 执行模型修复动作
      $t_1 = \pi_1(\sigma)$, $t_2 = \pi_2(\sigma)$
      if $| t_1 \cdot | = = 1$ and $| \cdot ((t_1 \cdot) \cdot) | > 1$
        p not in P
        $F' = F \cup \{((t_1, p) \in T \times \{p\}\} \cup$
        $\frac{\{(p, t_2) \in \{p\} \times T\}}{\{(t_1, t_1 \cdot) \in T \times P\}} \cup \{(t_1 \cdot, (t_1 \cdot) \cdot) \in P \times T\}$
        $P' = P \cup \{p\} / \{t_1 \cdot\} \cdot (t_2 \cdot) = \cdot (t_2 \cdot) \cup \cdot (t_1 \cdot)$
      else p not in P
        $F' = F \cup \{(\pi_1(\sigma), p) \in T \times \{p\}\} \cup$
        $\{p, (\pi_2(\sigma)) \in \{p\} \times T\}$
        $P' = P \cup \{p\}$
return $N_{repaired}$

算法 2 充分性证明:设网 $N = (P, T, F, M_0)$ 为有界 Petri 网,$RG(N) = (R(M_0), E, P)$ 为网 N 对应的可达标识图,日志 L 为事件轨迹的集合,$\exists t_a, t_b \in T$, $\forall M_i, M_j \in R(M_0)$,$\nexists M_k \in R(M_0) : M_i [t_a > M_k [t_b > M_j$,同时 $\forall \sigma \in L, \exists \lambda(t_a) \in L : \lambda(t_b) \in L$,那么一定存在变迁序列 $t_2, t_3, \cdots, t_{k-1}$ 和标识序列 $M_1, M_2, \cdots, M_k$,使得 $M_1 [t_b > M_2 [t_2 > M_3 \cdots M_{k-1} [t_a > M_k$.

算法 2 效率分析:上述算法遍历偏差集 R_d 迭代产生修复序列集 RS,通过遍历修复序列集 RS,仅保留有

效的修复序列到集合 RS' 执行模型修复动作. 保留有效修复序列即偏差约减过程. 假设存在修复序列 $<t_1,t_n>,<t_2,t_n>\cdots<t_{n-1},t_n>$,其中 $t_1\rightarrow t_2\rightarrow\ldots\rightarrow t_{n-1}$,基于偏差约减的模型修复方法基于修复序列添加新的次序关系 $t_{n-1}\rightarrow t_n$,提升了原始模型的精确度,同时保证了修复序列实现间接次序关系 $t_1\rightarrow t_2\rightarrow\ldots\rightarrow t_{n-1}\rightarrow t_n$,符合最终模型修复的目的. 相比于迭代的过程发现技术将所有的偏差全部进行修复,偏差约减的过程最大程度上保留了原有模型的结构,避免产生复杂的修复结果,所以该算法在执行过程中降低了模型修复的代价,减少了模型修复的时间. 本文在实验阶段通过大量的实验证明了该算法在精确度、拟合度、简洁度及时间复杂度方面有较好的表现.

由上一节可知网 N_1 与日志 L 的偏差集 $R_d=\{<t_4,[0,1,0,0,0,0,1,0],t_1>,<t_4,[0,0,0,1,0,0,1,0],t_3>,<t_1,[0,0,0,1,0,0,1,0],t_3>\}$,通过算法 2 遍历偏差集 R_d 保留与邻边关系差序列集 PS 中不同的元素得到修复序列集 $RS=\{<t_1,t_4>,<t_3,t_4>\}$. 在得到集合 RS' 的过程中,两个修复序列 $<t_1,t_4>,<t_3,t_4>$ 之间第二个元素相同,并且通过遍历日志邻边序列集 AS 可知序列第一个元素之间存在次序关系即 $t_1\rightarrow t_3$,保留序列 $<t_3,t_4>$ 作为修复序列,在模型修复的过程中,对于保留的序列之间执行模型修复动作时添加直接次序关系 $t_3\rightarrow t_4$,同时非保留的序列之间也达到了包含间接次序关系的效果即 $t_1\rightarrow t_3\rightarrow t_4$. 执行模型修复动作时,序列 $<t_3,t_4>$ 中 t_3 的后继变迁有多个前集库所 $|\cdot((t_3\cdot)\cdot)|>1$ 并且 t_3 的后集库所只有一个,此时首先删除 t_3 与后继变迁之间的流关系,然后在 t_3 与 t_4 之间添加直接次序关系并将 t_3 后继库所的前集变迁指向 t_4. 模型修复结果如图 3 所示. 从修复结果可以看出数据审批作为工具审批的条件之一,符合正确的数据加工流程.

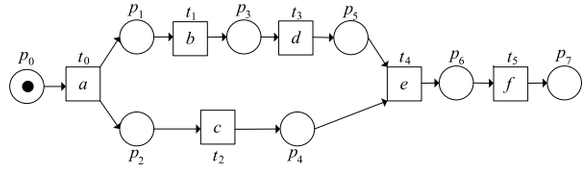

图3 修复模型

5 实验

本文实验采用的数据为天元数据网的大数据交易平台交易日志,仿真实验在 ProM6.0 平台上完成. 通过实验仿真与其他模型修复方法比较分析,进一步验证本文方法的正确性与有效性.

5.1 实验模型与数据

实验所使用的过程模型为天元数据网的大数据交易平台数据开发的过程模型,图展示交易平台数据开发的整个流程,表 2 为大数据交易过程中符号含义对照表.

表 2 符号含义对照表

符号	含义	符号	含义
T_1	用户申请	T_{13}	设置成员权限
T_2	用户申请审批	T_{14}	生成订单
T_3	进入数据中心	T_{15}	服务配置
T_4	项目申请	T_{16}	项目授权
T_5	选择数据表	T_{17}	订单审批
T_6	进入数据引擎	T_{18}	数据表授权加工
T_7	邀请项目成员	T_{19}	数据加工
T_8	申请数据表	T_{20}	导出数据
T_9	免费试用	T_{21}	部署
T_{10}	周期计费	T_{22}	人群投放
T_{11}	按次计费	T_{23}	应用开发
T_{12}	选择数据服务	T_{24}	上架售卖

据天元数据网的大数据交易平台交易日志,基于文献[13]中的 α 算法通过模型发现得到如图 4 所示的大数据交易流程数据应用开发模型作为原始模型.

实验过程中随机选取了 5 组事件日志 $L_1\sim L_5$ 用于模型修复,日志中都包含偏离模型的案例. 表 3 为事件日志统计表,表中每一列分别表示日志中案例总数,事件总数,所有轨迹的长度范围以及日志中轨迹偏离模型的个数.

表 3 事件日志统计表

Log	Total number of cases	Total number of events	Length of traces	Deviations
L_1	1048	21569	18 – 21	375
L_2	983	18629	14 – 19	256
L_3	1065	21025	16 – 20	326
L_4	1107	20872	13 – 19	297
L_5	925	19127	15 – 21	213

以日志 L_1 为例,按照基于偏差约减的模型修复方法对原始模型进行修复,通过在存在偏差的变迁之间添加约束的方式达到模型修复的目的. 图 5 为针对事件日志 L_1 执行基于偏差约减模型修复的结果,模型修复的过程中最小状态偏差率为 0.05. 从图 5 中可以看出,在多个并行分支之间,本文方法检测并修复了事件日志与模型之间的偏差,而且模型基本保持了原有模型的结构,模型修复的代价较小. 基于模型校准的模型修复方法在模型校准的过程中无法检测到多个并行分支之间事件与模型之间的偏差,因此原始模型经过模型修复之后的结果依旧是原先的模型,即图 4. 基于迭代

的过程发现技术对日志和模型中关系不同的地方根据关联规则产生新的关系. 关联规则为形如 $X \rightarrow Y$ 的蕴含式,X,Y 分别称为关联规则中的先导和后继,其中关联规则中的先导和后继分别用日志中的事件比表示,新的关系由日志或模型中支持度与置信度较高的决定. 由新的关系执行模型发现进行模型修复,修复结果如图 6 所示,从图 6 中可以看出修复过程中基本保留了所有的偏差关系,修复代价较大.

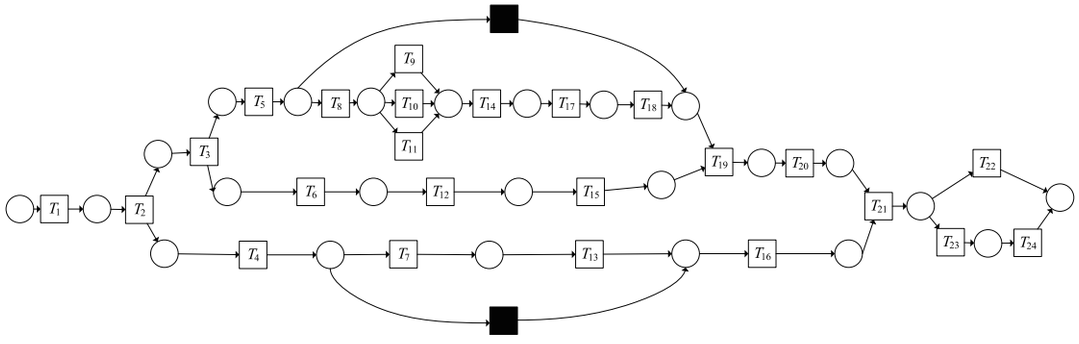

图4　大数据交易流程数据应用开发模型

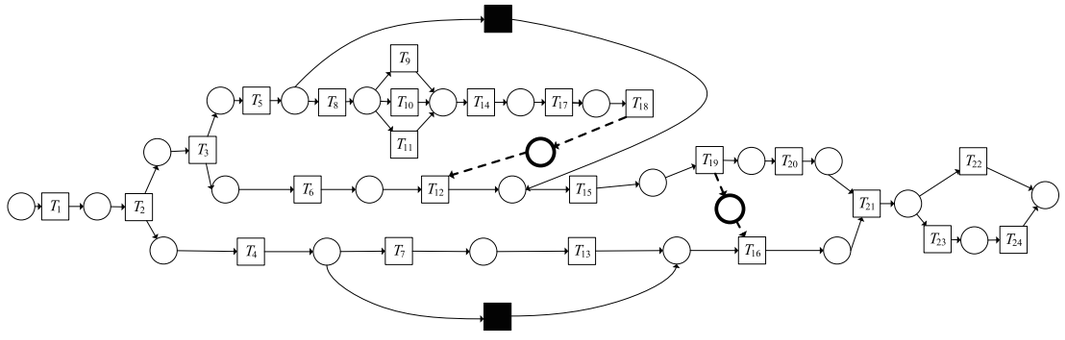

图5　基于L_1对图4的修复—基于偏差约减的模型修复

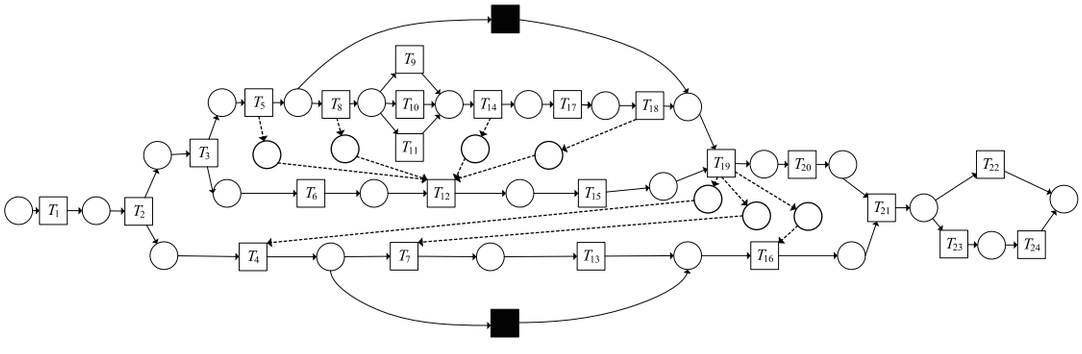

图6　基于L_1对图4的修复—基于迭代过程发现技术的模型修复

5.2　模型评估

本节对不同的模型修复方法产生修复结果从拟合度,精确度,简洁度以及时间复杂度几个角度进行对比分析.

首先分析不同方法产生的修复模型在不同数量级的事件日志中变化情况,模型评估的过程中拟合度与简洁度的计算采用文献[17]中的方法.

从图 7 和图 8 中可以看出在拟合度和简洁度方面,由于迭代的过程发现模型修复方法在模型中添加过多约束,导致日志重演的过程中不拟合的节点增多,同时也导致修复模型更加复杂,而其他两种方法基本保持原有模型的结构,所以基于偏差约减的模型修复和基于校准的模型修复的拟合度和简洁度基本一致且高于基于迭代过程发现的模型修复结果.

模型评估过程中日志与模型之间精确度的计算采用文献[18]中 conformance checker ++ 方法,从图 9 可看出在精确度方面,由于基于偏差约减的模型修复与基于迭代的过程发现技术的模型修复都在原有的模型

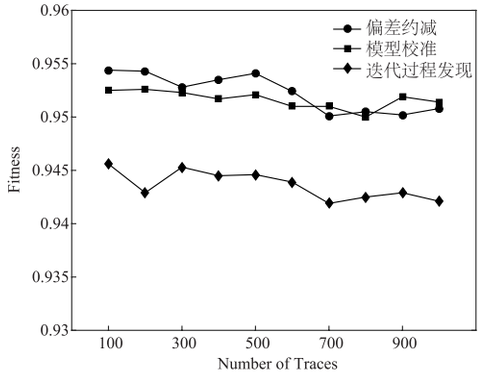

图7 日志–模型拟合度

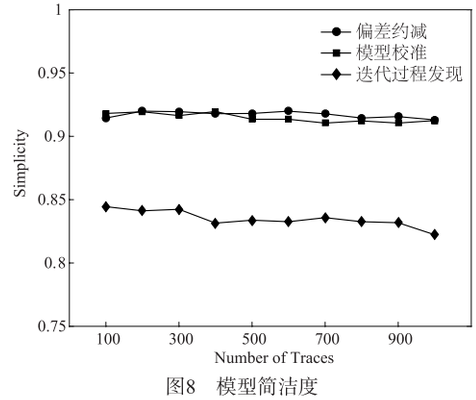

图8 模型简洁度

基础添加了约束条件,所以基于约减偏差关系的模型修复的精确度明显优于基于模型校准的模型修复的精确度,略低于基于迭代的过程发现技术的模型修复的精确度. 随着日志中轨迹的增加,不同的模型修复在精确度方法都呈现缓慢下降并最终平稳的趋势.

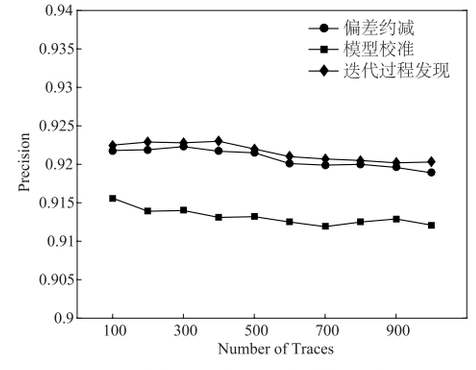

图9 日志–模型间的精确度

图 10 是 3 个模型修复方法分别对日志 $L_1 \sim L_5$ 修复所需的时间,由图 10 可知,基于偏差约减的模型修复所需的时间略低于基于校准的模型修复所需的时间,且明显低于基于迭代过程发现的模型修复所需的时间.

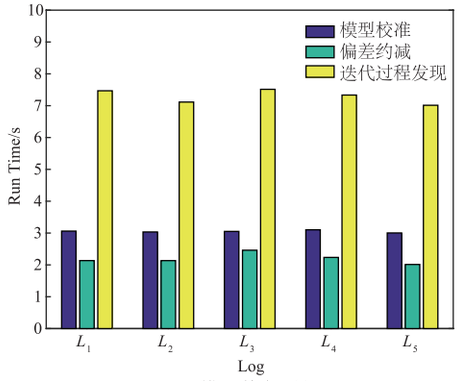

图10 模型修复时间

6 结论

本文通过分析大数据交易过程,发现如果模型修复过程中偏差出现在并行分支之间时,运用已有的模型修复方法无法检测出事件日志与模型之间的偏差,从而无法达到模型修复的目的. 针对上述问题,本文提出了基于偏差约减的模型修复方法能够准确的发现事件日志与模型之间的偏差,通过向存在偏差的变迁之间添加次序关系的方式执行模型修复动作,模型修复结果保持了原有模型的结构,大大的降低了模型修复的代价. 与此同时,通过与基于模型校准修复方法和基于迭代过程发现技术的模型修复方法的对比实验,从模型的拟合度、精确度、简洁度以及时间复杂度进行分析,验证了本论文方法的有效性与正确性.

本文的后续工作可以从以下三个方面继续开展研究:(1)在模型修复的过程中从多个角度分析日志,如时间角度或资源角度. (2)在模型修复的过程中考虑模型间行为一致性. (3)在模型修复的过程中考虑泛化度.

参考文献

[1] 余来文. 大数据商业模式[M]. 经济管理出版社,2014.

[2] 李骥宇. 大数据交易模式的探讨[J]. 移动通信,2016,40(5):41 – 44.

[3] 冯朝胜,秦志光,袁丁,等. 云计算环境下访问控制关键技术[J]. 电子学报,2015,43(2):312 – 319.
FENG Chaosheng, QIN Zhiguang, YUAN Ding, et al. Key techniques of access control for cloud computing[J]. Acta Electronica Sinica, 2015, 43(2): 312 – 319. (in Chinese)

[4] Aalst W M P V D, Hofstede A H M T, Weske M. Business process management: a survey[J]. Lecture Notes in Computer Science, 2008, 10(2): 1 – 12.

[5] Cheng Z, Zhu R, Chen P, et al. A distributed process management model for better scalability on multicore platform [J]. Chinese Journal of Electronics, 2017, 26(2): 263 – 270.

[6] 黄贻望,何克清,冯在文,等. 一种目标感知的可配置业务流程分析方法[J]. 电子学报,2014,42(10):2060-2068.
HUANG Yiwang,HE Keqing,et al. A goal-aware analytical method of configurable business process[J]. Acta Electronica Sinica,2014,42(10):2060-2068. (in Chinese)
[7] 于汪洋,黄昭,方贤文. 电子商务业务流程网的可达分析方法[J]. 电子学报,2017,45(7):1731-1739.
YU Wangyang,HUANG Zhao,FANG Xianwen. Reachability analysis methods of e-commerce business process net[J]. Acta Electronica Sinica,2017,45(7):1731-1739. (in Chinese)
[8] Aalst W M P V D,Medeiros A K A D,Weijters A J M M. Genetic process mining[J]. Lecture Notes in Computer Science,2006,14(2):76-83.
[9] van der Aalst W M P. Process Mining:Data Science in Action[M]. Springer,2016.
[10] 杜玉越,朱鸿儒,王路,等. 一种基于逻辑 Petri 网的过程挖掘方法[J]. 电子学报,2016,44(11):2742-2751.
DU Yuyue,ZHU Hongru,WANG Lu,et al. A method of process mining based on logic petri nets[J]. Acta Electronica Sinica,2016,44(11):2742-2751. (in Chinese)
[11] Pang S,Li Y,He H,et al. A model for dynamic business processes and process changes[J]. Chinese Journal of Electronics,2011,20(4):632-636.
[12] Adriansyah A,Munoz-Gama J,Carmona J,et al. Alignment Based Precision Checking[M]. Berlin Heidelberg:Springer,Business Process Management Workshops,2013. 137-149.
[13] Fahland D,Aalst W M P V D. Model repair-aligning process models to reality[J]. Information Systems,2015,47(1):220-243.
[14] Wynn M T,Wynn M T,Wynn M T,et al. Impact-driven process model repair[J]. ACM Transactions on Software Engineering & Methodology,2016,25(4):28-80.
[15] Aalst V D. Decomposing Petri nets for process mining:A generic approach[J]. Distributed & Parallel Databases,2013,31(4):471-507.
[16] Aalst W V D,Weijters T,Maruster L. Workflow Mining:Discovering Process Models from Event Logs[M]. IEEE Educational Activities Department,2004.
[17] Aalst W V D,Adriansyah A,Dongen B V. Replaying history on process models for conformance checking and performance analysis[J]. Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery,2012,2(2):182-192.
[18] 胡涛. 流程挖掘的一致性检查算法研究[D]. 华南理工大学,2015.
[19] Rozinat A,Aalst W M P V D. Conformance checking of processes based on monitoring real behavior[J]. Information Systems,2008,33(1):64-95.
[20] Mannhardt F,Leoni M D,Reijers H A,et al. Balanced multi-perspective checking of process conformance[J]. Computing,2016,98(4):407-437.
[21] Munoz-Gama J,Carmona J,Aalst W M P V D. Conformance checking in the large:partitioning and topology[A]. International Conference on Business Process Management[C]. Springer-Verlag,2013. 130-145.
[22] Adriansyah A,Dongen B F V,Aalst W M P V D. Towards Robust Conformance Checking[M]. Berlin Heidelberg:Springer,Business,Process Management Workshops,2010. 122-133.
[23] Sun W,Li T,Peng W,et al. Incremental workflow mining with optional patterns[A]. IEEE International Conference on Systems,Man and Cybernetics[C]. IEEE,2007. 2764-2771.
[24] Kalsing A C,Nascimento G S D,Iochpe C,et al. An incremental process mining approach to extract knowledge from legacy systems[A]. Enterprise Distributed Object Computing Conference[C]. IEEE,2010. 79-88.

作者简介

郭　艺　男,1993 年生于山东滨州,研究生,主要研究方向为过程挖掘、Petri 网、大数据.
E-mail:13646488081@163.com

叶　剑　男,1974 年生于山东济南,博士,高级工程师,硕士研究生导师. 主要研究方向为移动互联网挖掘、普适计算.
E-mail:jye@ict.ac.cn

张　鹏(通信作者)　男,1973 年生于山东泰安,山东科技大学计算机学院副教授,主要研究方向为 Petri 网、工作流、大数据、高性能计算等.
E-mail:bigbigroc@163.com