一种最优校准的分组算法

田银花[1,2],杜玉越[1]
(1. 山东科技大学 信息科学与工程学院,山东 青岛 266590;2. 山东科技大学 信息工程系,山东 泰安 271000)

摘 要:为了选取最优校准的代表项简化迹与模型的一致性检查,提出一种基于质数权值的分组算法,实现对所有最优校准的分组。给不同的偏差分配互异的质数权值,将权值之积作为最优校准的代价。包含相同移动多重集但移动出现位置不同的相似最优校准具有相同的代价。证明了分组算法的正确性,应用实例描述了分组的具体执行过程。算法首次明确简洁地实现了相似最优校准的分组,时间复杂度为 $O(m^2n^2)$。

关键词:迹;Petri 网模型;最优校准;质数权值;分组算法

中图分类号:TP311.13 **文献标志码**:A **文章编号**:1672-3767(2015)01-0029-06

A Grouping Algorithm of Optimal Alignments

Tian Yinhua[1,2], Du Yuyue[1]
(1. College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao, Shandong 266590, China;
2. Department of Information Engineering, Shandong University of Science and Technology, Taian, Shandong 271000, China)

Abstract: In order to select the representatives from optimal alignments and simplify the conformance checking between the trace and the model, a grouping algorithm based on prime weights was presented to group all the optimal alignments. The different prime weights were allocated to the deviations, and the products of the weights were taken as the costs of optimal alignments. Thus the similar optimal alignments with the same multiset of movements but different ordering shared the same cost. The correctness of this algorithm was proved and the specific implementation process was illustrated. The explicit and succinct grouping of similar optimal alignments was realized for the first time and the time complexity proves to be $O(m^2n^2)$.

Key words: trace; Petri net model; optimal alignment; prime weight; grouping algorithm

近年来,各企业组织普遍开始使用信息系统实施和管理业务,系统中存储和处理的数据以惊人的速度增长[1],同时,企业组织之间的竞争使得各组织必须优化运行模式,因而企业致力于对业务流程进行监控、跟踪和记录,以便提高服务质量。业务流程会在信息系统中留下足迹,形成事件日志。日志由大量独立事件组成,事件又称为迹(trace)[2]。日志反映了业务的实际运行情况,是分析业务流程、发现过程、处理偏差、改善流程以及进行决策支持的重要基础,因此,日志的记录与分析对于企业组织的发展与优化至关重要[3]。

企业组织一般根据其业务流程挖掘出相应的过程模型[4],与信息系统中所记录事件日志中的迹进行一致性检查(conformance checking)。若迹能在模型上重演,则该迹符合模型[5]。但是有些情况下,由于操作

收稿日期:2014-10-24

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:田银花(1982—),女,山东肥城人,讲师,博士研究生,主要从事流程挖掘和 Petri 网方面的研究. E-mail:skdxxtyh@163.com

杜玉越(1960—),男,山东聊城人,教授,博士生导师,CCF 高级会员,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

人员的失误、日志记录不准确或者模型本身不完善等原因，导致迹不能完全在模型上重演[6]。针对该情况，文献[7]给出一种衡量标准，并提出了最优校准(optimal alignment)的概念。

文献[7]给出的方法能够得到迹和 Petri 网[8]模型之间所有最优校准，并发现有些最优校准的移动多重集完全相同，只是移动出现的顺序不同，满足该条件的校准，具有相同的偏差，可以划分到同一组中，随机选择其中一个作为代表项进行分析即可。文献[7]给出了最优校准分组方法的描述，但未给出该方法的形式化定义，且根据该方法，很难实现最优校准的精确分组，甚至无法按照上述标准分组。

本研究的主要工作是在文献[7]的基础上，根据已有的 Petri 网模型，给定一条迹，得到迹与模型之间的所有最优校准，分析发现其相似性。提出一种基于质数权值的算法，能够实现此类相似最优校准的分组。该算法的时间复杂度为 $O(m^2n^2)$。

1 基本概念

本研究主要对 Petri 网模型与迹之间的最优校准进行分析，因此，首先给出标签 Petri 网[4]、迹和最优校准的相关概念。

定义 1(标签 Petri 网) 设 $A\subseteq\mathscr{A}$ 是一个活动标签的集合。集合 A 上的 Petri 网是一个元组 $N=(P,T,F,\alpha,m_i,m_f)$，其中：P 是库所的有限集合；T 是变迁的有限集合；$F\subseteq(P\times T)\cup(T\times P)$ 是有向弧的集合，称为流关系。$\alpha:T\to A^\tau$ 是一个标签函数。m_i,m_f 分别是 N 的初始标识和结束标识。τ 标记不可见变迁(invisible transition)。

当前信息系统记录了数量众多的事件，存储在日志中。日志中记录的一个事件，称为迹[9]。

定义 2(迹) 设 ε 为事件空间。$\sigma\in\varepsilon^*$ 是一个有限事件序列并且每个事件只出现一次，即对于 $1\leqslant i<j\leqslant|\sigma|:\sigma[i]\neq\sigma[j]$，则称 σ 为迹。

迹在 Petri 网模型上重演时，可能会出现偏差，校准能够标记偏差。给定对偏差的一个度量标准，得到最优校准[7]。

定义 3(校准) 设 $A\subseteq\mathscr{A}$ 是一个活动标签的集合。$\sigma\in A^*$ 是 A 上的一条迹，$N=(P,T,F,\alpha,m_i,m_f)$ 是 A 上的一个 Petri 网。迹 σ 与模型 N 之间的校准 $\gamma\in(A^{\gg}\times T^{\gg})$ 是满足以下条件的移动序列(movements)：

① $\pi_1(\gamma)_{\downarrow A}=\sigma$，即迹中的移动序列(忽略$\gg$)产生该迹；

② $m_i\xrightarrow{\pi_2(\gamma)_{\downarrow T}}m_f$，即模型中的移动序列(忽略$\gg$)产生一个完整的引发序列。

其中，$\gg$表示无移动(no move)，$A^{\gg}=A\cup\{\gg\}$，$\pi_1(\gamma)_{\downarrow A}$表示元组序列 γ 第 1 项在 A 上的投影。

对于校准中所有的元组$(a,t)\in\gamma$，对(a,t)的定义如下：

① 若 $a\in A$ 且 $t=\gg$，则为日志上的移动(move on log)；

② 若 $a=\gg$且 $t\in T$，则为模型上的移动(move on model)；

③ 若 $a\in A$ 且 $t\in T$，则为同步移动(synchronous move)；

④ 否则为非法移动(illegal move)。

$\Gamma_{\sigma,N}$记作迹σ 与模型 N 之间所有校准的集合。

定义 4(最优校准) 设 $A\subseteq\mathscr{A}$ 是一个活动标签的集合。$\sigma\in A^*$ 是 A 上的一条迹，$N=(P,T,F,\alpha,m_i,m_f)$ 是 A 上的一个 Petri 网。$l_c:A^{\gg}\times T^{\gg}\to IR$ 是移动的似然代价函数。称 $\gamma\in\Gamma_{\sigma,N}$ 为迹 σ 与模型 N 之间的最优校准，当且仅当对于任意的 $\gamma'\in\Gamma_{\sigma,N}$，使得 $\sum_{(a,t)\in\gamma}l_c(a,t)\leqslant\sum_{(a',t')\in\gamma'}l_c(a',t')$ 成立。

Γ^o_{σ,N,l_c} 记作迹 σ 与模型 N 之间基于似然代价函数 l_c 的所有最优校准的集合。

l_c 函数的取值直接决定了迹 σ 与模型 N 的最优校准集合，本文使用标准似然代价函数[7]。

定义 5(标准似然代价函数) 设 $A\subseteq\mathscr{A}$ 是一个活动标签集合。$N=(P,T,F,\alpha,m_i,m_f)$ 是 A 上的一个 Petri 网。标准似然代价函数 $l_c:A^{\gg}\times T^{\gg}\to IR$ 将所有移动映射到实数集上，对于任意$(x,y)\in A^{\gg}\times T^{\gg}$：

① $l_c((x,y))=0$，若 $x\in A,y\in T$ 且 $x=\alpha(y)$；或者 $x=\gg,y\in T$ 且 $\alpha(y)=\tau$；

② $l_c((x,y))=+\infty$，若 $x\in A, y\in T$ 且 $x\neq\alpha(y)$；或者 $x=y=>>$；

③ $l_c((x,y))=1$，其他情况。

2 相似最优校准

根据最优校准的定义可知，迹与模型之间的最优校准可能有多个，而且最优校准的共同点是包含的非同步移动个数相同，该共同点可以理解为最优校准之间的一种相似性。此外，对于最优校准之间的相似性还可以给出其他的度量标准，比如，最优校准中日志上的移动数目相同则认为它们是相似的。接下来，根据一个 Petri 网模型和一条迹，引出最优校准之间的另外一种相似性。

对文献[7]中已有模型进行适当改动，目的是去掉原模型中的重复变迁，但是模型的模拟能力没有改变，改动后的模型 N_{eb} 如图 1 所示。用户在电子书城进行在线交易，首先往购物车中添加货物(*add items*)，此过程可以反复执行，添加完后，若不想购买了，则放弃(*abort*)；若要购买的话，结束(*finalize*)添加过程，之后用户为购买的书籍付款(*pay*)，同时商家将书籍打包(*pack*)，然后系统进行有效性检查(*validate*)，检查无误商家将书籍投递出(*deliver*)，否则取消(*cancel*)本次交易。

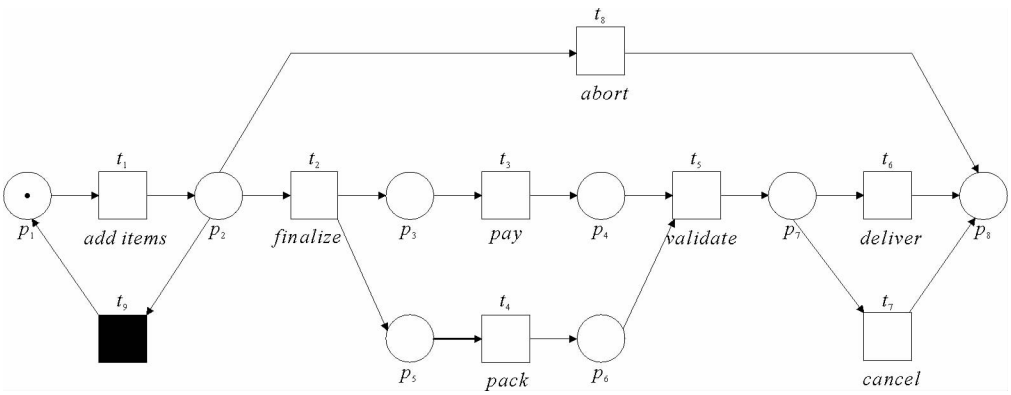

图 1 电子书城在线交易 Petri 网模型 N_{eb}

Fig. 1 The Petri net model N_{eb} of an online transaction in an electronic bookstore

过程实例在事件日志中记录为迹。给定迹和 Petri 网模型，如果迹中的每个活动都能由模型中的变迁引发模仿并且到达终止状态，则认为迹完全符合模型。反之，如果一条迹中不是所有的活动都能被模型中的变迁引发模仿，则认为迹不完全符合模型。迹 $\sigma_1=\langle add\ items, finalize, pay\rangle$ 是一条和模型 N_{eb} 存在偏差的迹，而且迹 σ_1 与模型 N_{eb} 之间存在多个校准。在对校准进行评价时，使用标准似然代价函数，得到迹 σ_1 与模型 N_{eb} 的所有最优校准如图 2 所示。在图 2 中，校准的上面一行表示迹 σ_1，下面一行表示模型 N_{eb} 的一个完整引发序列。无论是校准的上面一行还是下面一行，凡是发生偏差的位置都表示无移动，用符号>>来标记。

对最优校准进行分析，发现有些最优校准具有以下性质：包含的移动多重集完全相同，只是移动在多重集中出现的顺序不同。例如：将 γ_1 第 3 列上的移动(pay, t_3)与第 4 列上的移动$(>>, t_4)$交换位置可得到 γ_2；同样，γ_3 交换相同两列的内容得到 γ_4。除此之外，将 γ_5 第 2 列上的移动$(>>, t_8)$与第 3 列上的移动$(finalize, >>)$交换位置得到 γ_6，而 γ_6 第 3 列上的移动$(>>, t_8)$与第 4 列上的移动$(pay, >>)$交换位置得到 γ_7。即 γ_1 与 γ_2，γ_3 与 γ_4，γ_5 与 γ_6、γ_7 之间分别满足上述性质。

具有此类性质的校准称为相似最优校准。相似最优校准具有上述相似性，说明它们发生偏差的类型完全相同。因此，对模型和日志进行一致性检查时，将相似最优校准进行分组，只需对每组中的代表项进行研究即可，这样可以更好地对偏差进行诊断。

第 3 节是本研究提出的实现符合上述条件的相似最优校准分组的算法。

$\gamma_1=$

add items	finalize	pay	>>	>>	>>
add items	finalize	pay	pack	validate	deliver
t_1	t_2	t_3	t_4	t_5	t_6

$\gamma_2=$

add items	finalize	>>	pay	>>	>>
add items	finalize	pack	pay	validate	deliver
t_1	t_2	t_4	t_3	t_5	t_6

$\gamma_3=$

add items	finalize	pay	>>	>>	>>
add items	finalize	pay	pack	validate	cancel
t_1	t_2	t_3	t_4	t_5	t_7

$\gamma_4=$

add items	finalize	>>	pay	>>	>>
add items	finalize	pack	pay	validate	cancel
t_1	t_2	t_4	t_3	t_5	t_7

图 2　迹 σ_1 与模型 N_{eb} 的所有最优校准

Fig. 2　All optimal alignments between trace σ_1 and model N_{eb}

3　相似最优校准分组算法

根据文献[7]的算法 2 求解迹与模型之间所有最优校准，会生成一个变迁系统，表示事件网（由迹生成的网）与过程网（原 Petri 网）积的所有可达状态。其最优校准分组思想是选择该变迁系统中最优校准分支上的某个或者某些结点作为 knot 结点集，具有指向同一个 knot 结点的边所在路径产生的最优校准，被认为是相似最优校准。例如：所有路径都有边指向终点，若仅以终点为 knot 结点集，则所有最优校准均被认为是相似的，分到同一组中；若仅以源点为 knot 结点集，由于没有任何边指向源点，那么每个最优校准属于不同的组。

该算法思想本身比较灵活，选择的 knot 结点集不同，得到的最优校准分组不同；缺点是 knot 结点集的选择较为困难，很难选定恰当的 knot 结点集，实现符合要求且正确的相似最优校准的分组。

以第 2 节中迹 σ_1 与模型 N_{eb} 的所有最优校准进行分组为例。将所得变迁系统进行修剪，只保留包含最优校准所经过的边和结点，新得到的变迁系统有 12 个结点，任意选择一个或者多个结点作为 knot 结点集，knot 结点集的选择共 2^{12} 种组合，数量庞大。当然，在确定 knot 结点集的时候，可以根据结点在变迁系统中所处层次进行选择，即使如此，knot 结点集的选择也有 11 种情况，很难抉择选择哪些结点才能达到同组中最优校准之间包含的移动多重集相同而移动出现顺序不同的分组标准。甚至在该变迁系统中，不管选择哪些结点作为 knot 结点集都无法满足该要求，因此该算法很难实现上述相似最优校准的分组。

为解决上述问题，本研究给出一种相似最优校准分组算法。其主要思想是对所有最优校准进行扫描，若移动是同步的，则给它分配权值为 1；若移动是日志上的或者模型上的，则给它分配一个质数作为权值，且保证相同移动得到的权值相同，而不同移动得到的权值互异。然后，计算每个校准中所有移动权值之积作为校准的代价。具有相同代价的最优校准为相似最优校准，代价不同的最优校准不相似。

为便于描述，引入一些对序列的操作[9]。集合 A 上的任意序列 $\sigma=\langle a_1,a_2,\cdots,a_n\rangle$，$\partial_{set}(\sigma)=\{a_1,a_2,\cdots,a_n\}$ 且 $\partial_{multiset}(\sigma)=[a_1,a_2,\cdots,a_n]$。$\partial_{set}(\sigma)$ 将序列转换为集合，例如 $\partial_{set}(\langle d,a,a,a,a,a,a,d\rangle)=\{a,d\}$。$a$ 是 σ 的一个元素，写作 $a\in\sigma$，当且仅当 $a\in\partial_{set}(\sigma)$。$\partial_{multiset}$ 将序列转换为多重集，例如 $\partial_{multiset}(<d,a,a,a,a,a,a,d>)=[a^6,d^2]$。

另外，该算法要求模型中不存在重复变迁。符合该条件的模型必然能够得到以下结论：设 γ_1,γ_2 是迹 σ 与模型 N 之间两个最优校准，同步移动 $m_1,m_2\in\partial_{set}(\gamma_1)\cup\partial_{set}(\gamma_2)$，分别记作 $m_1=(a_1,t_1)$，$m_2=(a_2,t_2)$。若 $a_1=a_2$，必然 $t_1=t_2$，$m_1=m_2$ 成立，即该模型中不可能出现两个同步移动活动相同而变迁不同的情况。

给出该算法思想的形式化描述。首先，定义两个函数 $w(m)$ 和 $c(\gamma)$ 分别表示移动 m 的权值和最优校准 γ 的代价。其中，若 m 是同步移动，则 $w(m)=1$；否则，$w(m)=p$（p 为任意质数），且保证 m_1 和 m_2 是非同步移动时，若 $m_1\neq m_2$，则 $w(m_1)\neq w(m_2)$。$c(\gamma)=\prod_{i=1}^{|\gamma|}w(m_i)$。设 γ_1,γ_2 是迹 σ 与模型 N 之间两个最优校准。若 $c(\gamma_1)=c(\gamma_2)$，那么 γ_1,γ_2 是相似最优校准；反之亦然。定理 1 说明了算法思想的正确性。

为了方便定理的证明,定义以下多重集:

1)$B_{\text{all}}(\gamma)=\partial_{\text{multiset}}(\gamma)$表示一个包括最优校准 γ 中所有移动的多重集;

2)$B_{\log}(\gamma)=[m=(a,t)\in\partial_{\text{multiset}}(\gamma)\wedge t=>>]$表示一个包括 γ 中所有日志上的移动的多重集;

3)$B_{\text{mod}}(\gamma)=[m=(a,t)\in\partial_{\text{multiset}}(\gamma)\wedge a=>>]$表示一个包括 γ 中所有模型上的移动的多重集;

4)$B_{\text{syn}}(\gamma)=[m=(a,t)\in\partial_{\text{multiset}}(\gamma)\wedge(a\neq>>\wedge t\neq>>)]$表示一个包括 γ 中所有同步移动的多重集。

之所以给出多重集的概念,主要是因为 γ 是移动序列,序列中的元素是有顺序的,而多重集中的元素没有顺序。若最优校准 γ_1 与 γ_2 相似,$\gamma_1\neq\gamma_2$,但 $B_{\text{all}}(\gamma_1)=B_{\text{all}}(\gamma_2)$;反之亦然。另外,根据上述 4 个多重集的定义可得 $B_{\text{all}}(\gamma)=B_{\log}(\gamma)\uplus B_{\text{mod}}(\gamma)\uplus B_{\text{syn}}(\gamma)$。关于多重集的并运算$\uplus$,文献[7]中给出如下定义:$A$ 为集合,B_1,B_2,B_3 为 A 上的多重集。$B_3=B_1\uplus B_2$,当且仅当对于所有的元素 $a\in A$:$B_3(a)=B_1(a)+B_2(a)$。其中,$B_1(a)$标记元素 a 在多重集 B_1 中出现的次数。

定理 1 两个最优校准相似,当且仅当两个最优校准的代价相等。

证明:必要性。设 γ_1,γ_2 是迹 σ 与模型 N 之间的两个相似最优校准。根据上述对于相似最优校准性质的分析可知,γ_1 与 γ_2 中含有的移动多重集相同,只是移动出现顺序不同,因此 $c(\gamma_1)=\prod_{i=1}^{|\gamma_1|}w(m_i)=\prod_{j=1}^{|\gamma_2|}w(m_j)=c(\gamma_2)$。结论成立。

充分性。设 γ_1,γ_2 是迹 σ 与模型 N 的最优校准且 $c(\gamma_1)=c(\gamma_2)$。将 $c(\gamma_1)$和 $c(\gamma_2)$因式分解得到的结果必然相同,假设 $c(\gamma_1)=c(\gamma_2)=c_1\times c_2\times\cdots\times c_k$(其中 $c_i<c_j$,$1\leqslant i<j\leqslant k$ 且每一个因子都是质数不可继续分解)。因式中的每一个 c_i 对应一个权值。由 w 定义可知,当 $w\neq1$ 时,对应一个日志上或者模型上的移动,因此,$B_{\log}(\gamma_1)=B_{\log}(\gamma_2)$,$B_{\text{mod}}(\gamma_1)=B_{\text{mod}}(\gamma_2)$,$\pi_1(B_{\log}(\gamma_1))_{\downarrow A}=\pi_1(B_{\log}(\gamma_2))_{\downarrow A}$。记 $m=(a,t)$,对于任意的 $m\in B_{\text{mod}}(\gamma)$,因为 $a=>>$,所以 $\pi_1(B_{\text{mod}}(\gamma_1))_{\downarrow A}=\pi_1(B_{\text{mod}}(\gamma_2))_{\downarrow A}=\Phi$。又因为 $\pi_1(\gamma_1)_{\downarrow A}=\pi_1(\gamma_2)_{\downarrow A}=\sigma$,则 $\pi_1(B_{\text{all}}(\gamma_1))_{\downarrow A}=\pi_1(B_{\text{all}}(\gamma_2))_{\downarrow A}$,所以 $\pi_1(B_{\text{syn}}(\gamma_1))_{\downarrow A}=\pi_1(B_{\text{syn}}(\gamma_2))_{\downarrow A}$。模型中无重复变迁,$B_{\text{syn}}(\gamma_1)=B_{\text{syn}}(\gamma_2)$。因此,$B_{\text{all}}(\gamma_1)=B_{\log}(\gamma_1)\uplus B_{\text{mod}}(\gamma_1)\uplus B_{\text{syn}}(\gamma_1)=B_{\log}(\gamma_2)\uplus B_{\text{mod}}(\gamma_2)\uplus B_{\text{syn}}(\gamma_2)=B_{\text{all}}(\gamma_2)$。两最优校准相似得证。证毕。

算法 1 质数权值算法

输入:迹 σ 与模型 N 之间所有最优校准的集合 $\Gamma^{o}_{\sigma,N,l_c}=\{\gamma_1,\gamma_2,\cdots,\gamma_n\}$。

输出:$l[1,2,\cdots,n]$,其中 $l[i]$记录了与 γ_i 相似的最优校准的下标值中的最小者。

中间变量:$p[1,2,\cdots,nm]$存储质数;$c[1,2,\cdots,n]$记录每个最优校准的代价;$w[1,2,\cdots,n][1,2,\cdots,m]$记录最优校准中每个移动的权重值;$\gamma$ 记录最优校准中所有模型上的移动和日志上的移动。

Step 1 $p[1,2,\cdots,nm]=\{2,3,5,7,\cdots\}$;$l[1,2,\cdots,n]=\{1,2,\cdots,n\}$。

Step 2 遍历 γ_i,在遍历的过程中执行下述操作:

1)若 γ_i 中第 j 个移动为同步移动,置 $w[i][j]=1$;

2)若第 j 个移动不是同步移动,查看其是否在 γ 中,若在 γ 中且下标为 k,则置 $w[i][j]=p[k]$;

3)若不在 γ 中,若 γ 中有 h 个数据,将该移动存储于$(h+1)$位置,且置 $w[i][j]=p[h+1]$;

重复上述操作,直至 γ_i 中所有的移动都被遍历过。

Step 3 对于每个 γ_i,$c[i]=\prod_{i=1}^{|\gamma_i|}w[i][j]$;

Step 4 对于任意的 $i\neq j$,如果 $c[i]=c[j]$且 $i<j$,则置 $l[j]=l[i]$。

记最优校准的个数为$|\Gamma^{o}_{\sigma,N,l_c}|=n$,最优校准最大长度为 $\max(|\gamma_1|,|\gamma_2|,\cdots,|\gamma_n|)=m$。在算法中,需要存储所有的最优校准、每个移动的权值所在的二维数组以及每个最优校准代价的一维数组等,其中占用内存最多的为所有最优校准,最多包含 mn 个移动。因此算法的空间复杂度相对较为稳定,为 $O(mn)$。在最坏的情况下,所有最优校准中的移动都非同步,其个数为 nm。算法的重复执行步骤主要在 step 2,其重复执行次数为 $O(m^2n^2)$,因此算法的时间复杂度为 $O(m^2n^2)$。

利用上述算法,分析第 2 节中给出的实例。权值的分配和代价的计算分别如表 1 和表 2 所示。

表 1 权值分配表

Tab. 1 Weight allocation table

移动	$(add\ items, t_1)$	$(finalize, t_2)$	(pay, t_3)	$(>>, t_4)$	$(>>, t_5)$	$(>>, t_6)$	$(>>, t_7)$	$(>>, t_8)$	$(finalize, >>)$	$(pay, >>)$
权值	1	1	1	2	3	5	7	11	13	17

表 2 代价表

Tab. 2 Cost table

最优校准	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7
代价	$2\times3\times5=30$	$2\times3\times5=30$	$2\times3\times7=42$	$2\times3\times7=42$	$11\times13\times17=2\ 431$	$13\times11\times17=2\ 431$	$13\times17\times11=2\ 431$

通过表 1 可以看出,同步移动的权值均为 1,非同步移动的权值分别为不同的质数。接下来,根据表 1 中各个移动的权值,计算每个最优校准的代价。计算方法为将最优校准中非同步移动的权值相乘,所得结果即为该最优校准的代价。根据定理 1 可知,代价相同的最优校准是相似的,因此,该实例中 γ_1 与 γ_2,γ_3 与 γ_4,γ_5、γ_6 与 γ_7 分别为相似最优校准。

4 结束语

通过实例详细阐述了最优校准之间的相似性,分析了现有最优校准分组的算法思想,发现该方法虽然灵活但是很难真正实现此类相似最优校准分组。因此,提出质数权值算法,其主要思想是为出现在最优校准中的移动分配权值,原则是同步移动的权值为 1,非同步移动的权值为质数,且保证相同的非同步移动权值相同,不同的非同步移动权值互异;将每个最优校准中的所有移动权值相乘,作为该最优校准的代价,代价相同的最优校准是相似最优校准。此算法适用于模型中无重复变迁的情况。通过定理证明了算法的正确性,确保最优校准之间只要代价相同,则两个最优校准包含的移动多重集相同,只是移动出现的位置不同。算法的空间复杂度为 $O(mn)$,时间复杂度为 $O(m^2n^2)$。

参考文献:

[1]Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition and productivity[EB/OL]. [2011-07-01][2014-10-20]http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

[2]Buijs J. Mapping data sources to XES in a generic way[D]. Eindhoven, the Netherlands: Eindhoven University of Technology, 2010: 21-59.

[3]李军. 大数据:从海量到精准[M]. 北京:清华大学出版社, 2014: 4-26.

[4]van der Aalst W M P, Stahl C. Modeling business processes: A Petri net oriented approach[M]. Cambridge, MA: MIT Press, 2011: 13-75.

[5]Bose R P J C, van der Aalst W M P. Process diagnostics using trace alignment: Opportunities, issues and challenges[J]. Information System, 2012, 37(2): 117-141.

[6]van der Aalst W M P, Adriansyah A, Dongen B F. Replaying history on process models for conformance checking and performance analysis[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2012, 2(2): 182-192.

[7]Adriansyah A. Aligning observed and modeled behavior[D]. Eindhoven, the Netherlands: Eindhoven University of Technology, 2014: 35-97.

[8]Murata T. Petri nets: Properties, analysis and applications[J]. Proceedings of the IEEE, 2002, 77(4): 541-580.

[9]van der Aalst W M P. Process mining: Discovery, conformance and enhancement of business processes[M]. Berlin: Springer-Verlag, 2011: 75-123.

(责任编辑:吕文红)