

# 基于成本对齐的业务流程变化挖掘方法



刘 静 方贤文

安徽理工大学数学与大数据学院 安徽 淮南 232001

(1825852939@qq.com)

**摘 要** 变化挖掘是业务流程管理的核心,从事件日志中挖掘出业务流程的变化尤为重要。已有对变化挖掘的分析方法大多集中在源模型或目标模型已知的基础上。文中从系统日志的角度提出了一种基于成本最优对齐的业务流程变化挖掘方法。首先,根据事件日志提取出有效的高频形态学发生片段,计算出各迹对齐时的最高成本函数值,并在此基础上发现最优迹对齐;然后,通过度量最优对齐时变化日志与源日志间的相似性度,快速且高效地挖掘出变化集。最后通过实例分析显示了该方法的有效性。

**关键词:** 变化挖掘;最优对齐;成本函数;相似性度;高频片段

**中图法分类号** TP391.9

## Mining Method of Business Process Change Based on Cost Alignment

LIU Jing and FANG Xian-wen

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

**Abstract** Change mining is the core of business process management, and it is particularly important to mine the changes of business processes from the event log. Most of the existing analysis methods of change mining focus on the source model or target model. From the point of view of system log, this paper proposes a business process change mining method based on cost optimal alignment. Firstly, according to the event log, the effective high frequency morphological occurrence segment is extracted, the highest cost function value of each trace alignment is calculated, and on this basis, the optimal trace alignment is found, and then the similarity between the change log and the source log is measured to mine the change set quickly and efficiently. Finally, an example is given to show the effectiveness of the method.

**Keywords** Change mining, Optimal alignment, Cost function, Similarity degree, High frequency segment

### 1 引言

随着大数据技术的快速发展,业务流程管理及优化在企业或组织中扮演着越来越重要的角色。然而,建模时一个小的细节发生错误,或者对流程施加外部更改(例如新的法规或公司指南)时,用户将被迫偏离规定的业务流程模型。企业在产品开发过程中更好地管理工程变更(Manage Engineering Changes, MEC)的能力可以降低成本,缩短开发时间,并生产出更高质量的产品。同时,由于重新设计流程是一项昂贵且耗时的任务,因此正确、有效地识别和挖掘业务流程变化在流程管理及优化方面具有较高的应用价值。

目前国内外对业务流程变化的研究主要集中在变化域、变化传播和变化挖掘上。业务流程变化域的研究主要集中在

通过源模型和目标模型之间的对应关系和动态切片技术精确定位变化域和变化域融合分析上<sup>[1-3]</sup>。而业务流程变化传播的分析则在变化域研究的基础上,更深层次地探究模型中的变化对其他相关模型产生的影响<sup>[4-7]</sup>。文献[8-9]集成了一个框架来支持 EA 模型中的变化传播,用于从约束 EA 模型的一致性规则生成交互式修复计划,在业务设置发展时维护业务服务与其支持的 IT 资产之间的一致性。除了变化域和变化传播方面的研究,业务流程的变化挖掘也成为了近年来的热点课题<sup>[10]</sup>。文献[11]提出了一种在自适应过程管理系统中挖掘变化日志的方法,通过流程挖掘发现的更改流程提供了到目前为止发生的所有变化的聚合概述,它可能导致建模者重新设计流程模型。文献[12]讨论了从流程模型适应中学习和发现一个参考模型的两种场景,其中可以用最小的成本

到稿日期:2019-06-26 返修日期:2019-10-02 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。

基金项目:国家自然科学基金(61272153,61402011);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)

This work was supported by the National Natural Science Foundation of China(61272153,61402011), Natural Science Foundation of Anhui Province(1508085MF111,1608085QF149), and Natural Science Foundation of Colleges and Universities in Anhui Province(KJ2016A208).

通信作者:方贤文(280060673@qq.com)

配置变量将这些变体合并到一个新的参考流程模型中。文献[13]提出了一种启发式搜索算法,通过挖掘过程变量促进从过去的过程变化中学习,不仅可以控制更新引用模型的工作量,还可以获得最重要的适应性和灵活性。文献[14]提出的一种基于迹聚类的方法将事件日志划分为同构子集,并为每个子集创建流程模型,可以在实际的灵活环境中改进流程挖掘结果。文献[15]讨论了用于配置和管理业务流程变体的高级概念,它提供了一种灵活而强大的解决方案用于管理业务流程变体的生命周期,这种变体支持将促进变更系统的流程配置和流程维护。

这些研究方法大多是从业务流程参考模型已知的角度出发,很少基于事件日志的角度来挖掘变化。文献[16]从不完备日志的角度,基于日志的联合发生关系来挖掘系统的行为变化。该方法挖掘的是系统中单个或多个活动的行为关系的变化;而采用本文方法能得到发生变化的整条迹,更具完整性。本文在日志的基础上引入高频发生片段和成本函数,从最优对齐相似性度量的角度进行变化挖掘。

本文第2节给出一组系统记录到的变化前的源日志和变化后的实际日志的实例;第3节介绍相关定义;第4节首先基于日志挖掘出高频形态学片段,然后通过计算最优对齐得到的变化日志中的迹与源日志的相似性来挖掘发生变化的迹集;第5节针对第2节的动机例子进行实例分析;最后总结全文并展望未来。

2 动机例子

从系统中记录到的一组事件日志,称为源日志  $L_{or}$ ,如表1所列。

表1 源日志  $L_{or}$

Table 1 Source log  $L_{or}$

| 实例 | 事件轨迹    | 实例数 |
|----|---------|-----|
| 1  | HABGDF  | 5   |
| 2  | HACBDFG | 375 |
| 3  | HABCDGF | 462 |
| 4  | HABCDGF | 219 |
| 5  | HABDFG  | 288 |
| 6  | HABDGF  | 190 |
| 7  | HACBDFG | 23  |
| 8  | HABCDGF | 61  |

随着时间的演变,系统中某些行为可能会发生变化。表2列出了一组发生变化后系统所产生的实际日志,被称为变化日志  $L_{ch}$ 。本文中用到的日志均由工作流网系统产生,理论上默认为完备日志。

表2 变化日志  $L_{ch}$

Table 2 Change log  $L_{ch}$

| 实例 | 事件轨迹     | 实例数 |
|----|----------|-----|
| 1  | EHACBDFG | 448 |
| 2  | HABCDGF  | 437 |
| 3  | HEABCDGF | 400 |
| 4  | HEABDGF  | 310 |
| 5  | HABCDGF  | 27  |
| 6  | HEABCDGF | 57  |
| 7  | HABCDGF  | 125 |

为了精确挖掘变化,往往采用对齐的方法。已有研究在计算最优对齐时,引用文献[17]中的成本函数:  $Score(T_1, T_2) = \sum_{j=1}^m e_j$ 。其中:

$$e_j = \begin{cases} S(a,b), & \text{if } T_1(j)=a, T_2(j)=b \\ I(a,b), & \begin{cases} \text{if } T_1(j)=b, T_1(j-1)=a, \\ T_2(j)=- \\ \text{if } T_1(j)=-, T_2(j)=b, \\ T_2(j-1)=a \end{cases} \end{cases}$$

其中,  $T_1$  和  $T_2$  为日志中两条不相同的迹,  $S(a,b)$  为代替操作的成本值,  $I(a,b)$  为删除或插入操作的成本值。图1给出迹  $T_1=abcbac$  和  $T_2=acacac$  的一种对齐方式。

|       |   |   |   |   |   |   |
|-------|---|---|---|---|---|---|
| $T_1$ | a | b | c | a | c | - |
| $T_2$ | a | c | a | c | a | d |

图1  $T_1$  和  $T_2$  的对齐示例1

Fig. 1 Example 1 of alignment of  $T_1$  and  $T_2$

针对图1中的对齐示例,计算  $Score(T_1, T_2)$ ,  $\forall (a,b) \in \Sigma$ , 当  $a=b$  时,令  $S(a,b)=1$ ; 当  $a \neq b$  时,令  $S(a,b)=-1$ ,  $I(a,b)=-1$ 。由此计算出  $Score(T_1, T_2)_1=-4$ 。图2给出  $T_1=abcbac$  和  $T_2=acacac$  的第二种对齐方式( $T_1$  和  $T_2$  的对齐方式有多种,此处仅以这两种为例进行分析)。

|       |     |     |     |     |     |     |     |     |     |     |     |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $T_1$ | $a$ | $b$ | $c$ | $a$ | $c$ | $-$ | $-$ | $-$ | $-$ | $-$ | $-$ |
| $T_2$ | $-$ | $-$ | $-$ | $-$ | $-$ | $a$ | $c$ | $a$ | $c$ | $a$ | $d$ |

图2  $T_1$  和  $T_2$  的对齐示例2

Fig. 2 Example 2 of alignment of  $T_1$  and  $T_2$

采用文献[17]的方法分别计算上述两种对齐示例的成本函数值,可以发现当对  $S(a,b)$  和  $I(a,b)$  赋不同值时,迹对齐的成本函数值  $Score(T_1, T_2)$  也会不同。由于无法确定  $Score(T_1, T_2)$  值,因此无法计算出最优对齐。

从上述分析可以看出,成本函数  $Score$  值不同时最优对齐也会有所不同,因此有必要找到最大成本函数值。本文提出基于高频形态学发生片段的理论来计算最大成本函数值,从而得到最优对齐,进而基于最优对齐进行相似性度量,实现变化挖掘。

3 基本定义

定义 1<sup>[18]</sup>(工作流网)  $N=(S,T,F)$  为一个 Petri 网,当  $N$  为工作流网(WF-net)时,须符合下列条件:

- (1) 有且只有唯一源库所  $i: i \in \Phi$ ;
- (2) 有且只有唯一终止库所  $o: o \in \Phi$ ;
- (3)  $N'=(S, T \cup \{t_c\}, F \cup \{(o, t_c), (t_c, i)\})$  为网  $N$  的外延子网,且具有较强的连通性。

定义 2<sup>[19]</sup>(事件日志) 设  $\Sigma$  是活动变迁集合,  $\sigma \in \Sigma^*$  是一条事件轨迹,  $L \in B(\Sigma^*)$  是一个事件日志,其中  $B(\Sigma^*)$  是迹的多集。

定义 3<sup>[19]</sup>(日志行为轮廓) 设  $L_p=n_1, \dots, n_m$  为流程模型  $P=(A, a_1, a_0, C, F, T)$  中的一条日志变迁对  $(x, y) \in (A_L \times A_L)$  至多存在下面两种关系之一:

- (1) 严格序关系  $\rightarrow_L$ , 当且仅当  $x \succ_L y, y \not\prec_L x$ ;

(2)交叉序关系  $\parallel_L$ , 当且仅当  $x \succ_L y, y \succ_L x$ 。

**定义 4**<sup>[17]</sup>(迹对齐) 给定一组迹  $T = \{T_1, T_2, \dots, T_n\}$  和另一组迹  $\bar{T} = \{\bar{T}_1, \bar{T}_2, \dots, \bar{T}_n\}$ , 迹对齐则为  $T$  在  $\bar{T}$  上的映射,  $\bar{T}_i \in (\Sigma \cup \{-\})^+, 1 \leq i \leq n$ , 并且满足:

- (1)  $|\bar{T}_1| = |\bar{T}_2| = \dots = |\bar{T}_n| = m, m \in N$ ;
- (2) 在移动若干个“—”后,  $|T_i| = |\bar{T}_i|$ ;
- (3) 不存在  $k \in \{1, 2, \dots, m\}, \forall 1 \leq i \leq n, \bar{T}_i(k) = -$ 。

其中,  $m$  为迹  $\bar{T}$  的长度,  $l_{\max} \leq m \leq l_{\sum}$  且  $l_{\max}$  为  $T$  中最长迹的长度,  $l_{\sum}$  为  $T$  中所有迹的长度总和;  $\Sigma$  为活动集,  $\Sigma^+$  表示  $\Sigma$  的所有非空有限序列; “—”表示一个缺口。

**定义 5**<sup>[17]</sup>(迹对齐的操作方式) 给定迹  $T_1$  和  $T_2$ , 一对迹的对齐可以认为是  $T_1$  在  $T_2$  上的映射, 则对于迹对齐中的任意列有以下几种操作方式:

- (1) 活动对  $(a, b), a, b \in \Sigma$ , 表示  $T_1$  中的活动  $a$  由  $T_2$  中的活动  $b$  代替(sub 操作);
- (2) 活动对  $(a, -)$  表示活动  $a$  在  $T_2$  中删除;
- (3) 活动对  $(-, b)$  表示活动  $b$  在  $T_1$  中插入。

值得注意的是, 插入和删除操作是互补的, 因为一个迹中的插入可以被视为另一个迹中的删除。因此, 我们将插入和删除操作称为 indel 操作。

## 4 基于成本最优对齐的业务流程变化挖掘方法分析

本节主要介绍基于变化前后的事件日志之间的相似性度发现业务流程变化的位置, 并提出新的算法。该算法不仅解决了当前变化挖掘过程中只注重从模型角度分析的问题, 还基于变化日志中迹与源日志间的相似性度有效地挖掘出了完整的发生变化的迹, 使其在模型修复中具备更好的适用性。首先, 从给定的变化日志中提取出高频形态学发生片段, 并根据成本函数计算出最大  $Score$  值, 找到最优对齐; 然后, 基于最优对齐度量相似性大小, 并将度量结果与阈值进行比较, 从而发现变化迹集。

### 4.1 基于事件日志挖掘高频形态学片段

从形态结构的角度, 判定频繁性较高的过程片段即为高频形态学片段。基于事件日志挖掘出该类片段可以有效地计算最大成本函数值, 同时大大降低了最优对齐的复杂程度。本节提出了基于事件日志提取高频形态学发生片段的算法(见算法 1), 该算法通过给定的事件日志计算长度为  $l$  的片段的价值  $value$ , 并针对价值相等的片段计算全局频率, 从而挖掘高频形态学片段来研究最优对齐。下面给出过程片段和等价形态学片段的概念, 并执行算法 1, 以快速找出高频片段。

#### 算法 1 频繁等价的形态学片段提取方法

输入: 事件日志  $L$ (迹的多集)

输出: 等价形态学片段集合

Step1 对于数据集  $DS$ , 提取不同的迹  $t_i (1 \leq i \leq n)$ , 计算发生频率  $LF_{t_i}$ , 执行 Step2;

Step2 if  $\forall t_i \in DS$ ,  
 $LF_{t_i} < \beta (\beta = 0.05)$ , delete  $t_i$ ,  
 else if reserved  $t_i$ , 执行 Step3;

Step3 return  $\forall t_i \in DS$ , 产生所有长度为  $l$  的片段  $LG_{t_i} (3 \leq l \leq |t_{\max}|)$ ,

执行 Step4;

Step4  $\forall l\text{-gram} \in LG_{t_i}$ , 计算价值  $value(l\text{-gram}) = \omega \times A_1 + \sum_{j \neq 1, j \neq 0} A_j + \lambda \times A_0$  ( $\omega$  和  $\lambda$  分别是第一个活动和最后一个活动的系数, 并且是不同于其他活动系数的互异数值), 并判断  $value(l\text{-gram})$  是否相等, 执行 Step5;

Step5 if  $l\text{-gram}_1 = l\text{-gram}_2$ , 计算全局频率  $GF_1$ , 执行 Step6,  
 else if 重新执行 Step4;

Step6 if  $GF_1 > 1$ , return  $l\text{-gram}$ ,  
 else if delete  $l\text{-gram}$ , 执行 Step7;

Step7 返回所有  $GF_1 > 1$  的等价形态学片段, 执行 Step8;

Step8 算法结束。

该算法有效地从事件日志中发现了频繁的等价形态学片段, 为迹对齐的最优判别提供了可靠的依据。

**定义 6**<sup>[20]</sup>(过程片段) 对于一个过程片段  $F(A, G, R, s, e)$ ,  $F$  当且仅当满足下列条件时被定义为过程片段:

- (1)  $R \subseteq (A \times G) \cup (G \times A) \cup (G \times G)$ ;
- (2)  $|A| \geq 1$ ;
- (3)  $\forall t \in (A \cup G), \exists v_0 = s, v_1, v_2, \dots, v_k = t ((v_{i-1}, v_k) \in R, 1 \leq i \leq k)$ ;
- (4) if  $\exists n_1, n_2 (n_1 = s, n_2 = s)$ , then  $n_1 = n_2$ ;
- (5) if  $\exists n_1, n_2 (n_1 = e, n_2 = e)$ , then  $n_1 = n_2$ ;
- (6)  $\exists t \in (A \cup G), ((v_i, v_i) \in R \cup (v_k, v_i) \in R) \cup ((v_j, v_{j+1}) \in R, i \leq j \leq 1)$ 。

其中,  $A$  代表活动集合,  $G$  代表网关,  $R$  是控制流关系集合,  $s$  是单一的输入节点集,  $e$  是单一的输出节点集。

**定义 7**<sup>[20]</sup>(等价形态学片段) 若两个片段  $f_1(A_{f_1}, G_{f_1}, R_{f_1}, s_{f_1}, e_{f_1})$  与  $f_2(A_{f_2}, G_{f_2}, R_{f_2}, s_{f_2}, e_{f_2})$  以不同的方式执行相同的任务, 被称为等价的形态学片段, 当且仅当:

$$s_{f_1} = s_{f_2}, e_{f_1} = e_{f_2}, A_{f_1} = A_{f_2}$$

在定义 7 中, 两个片段有着相同的开始和结束节点 ( $s_{f_1} = s_{f_2}, e_{f_1} = e_{f_2}$ ), 中间节点集相同 ( $A_{f_1} = A_{f_2}$ ), 但是它们有不同的顺序和不同的关系。下面通过举例对过程片段及等价形态学片段进行说明, 如图 3 所示。

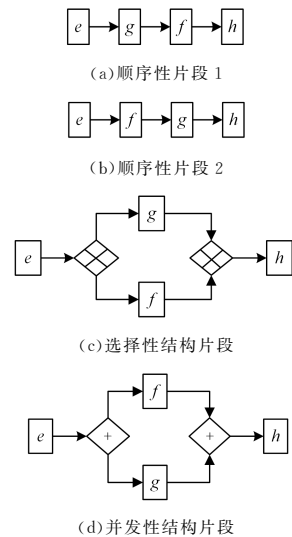


图 3 4 种等价形态学片段

Fig. 3 Four equivalent morphological fragments

图 3 给出了 4 种结构不同而操作相同的片段,它们分别是 4 种结构不同的过程片段,此类片段的特征为开始节点和结束节点相同,并且具有相同的中间节点。因此,图 3 所示 4 种片段被称为等价的形态学片段。

4.2 基于最优对齐的业务流程变化挖掘方法

业务流程在实际运行中可能会出现变化,而分析并找到这种变化是业务流程管理的重要内容,因为重新设计新的流程模型将会耗费较高的人力和物力成本。上文已发现了日志中有有效的高频发生活动,基于此,本节提出通过最优对齐的相似性度量挖掘系统中变化迹集的算法。该算法根据日志相似性度的知识计算迹与源日志的相似性度,并给定相似性度的阈值,满足阈值条件的迹,即为系统中发生变化的迹。

定义 8(相似性) 给定事件日志  $L_1$  和  $L_2$  为迹的多集,且  $L_1 = \{\sigma_1, \sigma_2, \dots, \sigma_m\}$ ,  $L_2 = \{t_1, t_2, \dots, t_n\}$ ,  $L_1$  和  $L_2$  之间的相似性度被定义为:

$$Sim(L_1, L_2) = \frac{f_{cost}(L_1, L_2)}{|L_1| \times move_{L_2}(L_2) + |L_2| \times move_{L_1}(L_1)}$$

其中,  $f_{cost}(L_1, L_2) = \sum_{\sigma_i \in L_1, t_j \in L_2} Score(\sigma_i, t_j)$  ( $1 \leq i \leq m, 1 \leq j \leq n$ ), indel 操作和 sub 操作统称为 move 操作。

本文在相似性度的基础上提出基于事件日志定位变化迹的算法,如算法 2 所示。

算法 2 迹变化挖掘算法

输入:源日志  $L_{or}$ ,变化日志  $L_{ch}$

输出:变化迹集  $\{t_{ch}\}$

- Step1 根据算法 1,计算  $L_{ch}$  中的所有高频片段 l-gram( $l=4$ ),执行 Step2;
- Step2  $\forall \sigma_i \in L_{or}, \forall t_j \in L_{ch} (1 \leq i \leq m, 1 \leq j \leq n)$ , 计算  $\sigma_i$  和  $t_j$  对齐时的成本函数  $Score(t_j, \sigma_i) = \sum_{t_j \in L_{ch}, \sigma_i \in L_{or}} e_j$ , 执行 Step3;
- Step3 if  $\exists Score'$ , 使得  $Score'(t_j, \sigma_i) > Score(t_j, \sigma_i)$ , 执行 Step4, else if 重新执行 Step2;
- Step4 return  $\forall t_j \in L_{ch}$ , 计算  $f_{cost}(t_j, L_{or}) = \sum_{j=1}^n Score(t_j, L_{or})$ , 执行 Step5;
- Step5  $\forall t_j \in L_{ch}$ , 计算所有的  $Sim(t_j, L_{or}) = \frac{f_{cost}(t_j, L_{or})}{|t_j| \times |\sigma_i|}$ , 执行 Step6;
- Step6 if  $\forall t_j \in L_{ch}, Sim(t_j, L_{or}) > \epsilon (\epsilon = 0.8)$ , deletet $t_j$ , else if  $Sim(t_j, L_{or}) < \epsilon$ , return $t_j$ , 执行 Step7;
- Step7 返回所有  $Sim(t_j, L_{or}) < \epsilon$  的迹  $t_j$ , 执行 Step8;
- Step8 算法结束。

5 实例分析

本节以第 2 节给定的源日志和变化日志(动机例子)为例来显示所提算法的可行性。针对算法 1 进行实例分析,考虑表 2 中的变化日志  $L_{ch}$ 。首先,对  $L_{ch}$  中的迹进行预处理,把相同轨迹的实例整合到一条迹中;然后,计算每一条迹的发生频率  $LF_{t_i}$ ,设定阈值  $\beta=0.05$ (因为是判断日志中的低频行为,且给定事件的日志实例数较少);对发生频率小于阈值  $\beta$  ( $LF_{t_i} < \beta$ ) 的迹进行过滤处理,保留在事件日志中发生频率较高的迹。由表 1 中的数据信息很容易计算每条迹的发生频率  $LF_{t_i}$ :

$$LF(EHACBDFG) = \frac{448}{1804} = 0.2483$$

$$LF(HABCDGFG) = \frac{589}{1804} = 0.3265$$

$$LF(HEABCDGFG) = \frac{457}{1804} = 0.2533$$

$$LF(HEABDGF) = \frac{310}{1804} = 0.1718$$

基于此,预处理后的日志  $L_{ch}'$  如表 3 所列。

表 3 预处理后的变化日志  $L_{ch}'$   
Table 3 Pre-processed change log  $L_{ch}'$

| 实例 | 迹 $t_i (1 \leq i \leq 4)$ | 实例数 |
|----|---------------------------|-----|
| 1  | HEABDGF                   | 310 |
| 2  | HABCDGFG                  | 589 |
| 3  | HEABCDGFG                 | 457 |
| 4  | EHACBDFG                  | 448 |

$L_{ch}'$  中的每一条迹  $t_i (1 \leq i \leq 4)$  都可以产生长度为  $l$  的片段  $LG_{t_i} (3 \leq l \leq |t_{max}|)$ ,  $l$  的最小值取 3,最大值为迹  $t_i$  中最长迹的长度。对于表 2 中的 4 条迹,其产生长度为  $l=4$  的片段,如图 4 所示。

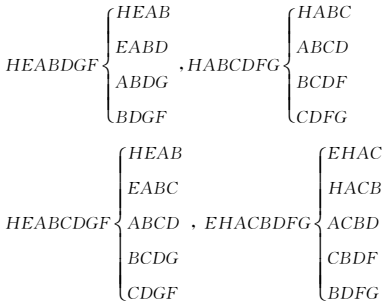


图 4 迹  $t_i$  产生的长度为  $l=4$  的片段  
Fig. 4 Fragment of length  $l=4$  produced by trace  $t_i$

在计算  $value(l\text{-}gram)$  时,为第一个和最后一个活动添加不同的系数  $\omega, \lambda (2 \leq \omega, \lambda \leq 9)$ ,它们的作用主要是区分第一个活动和最后一个活动产生的不同影响。假定  $\omega=2, \lambda=3$ ,首先将事件日志  $L_{ch}'$  中的所有活动映射为不同的数值,用 10 的幂次表示,如表 4 所列。

表 4 活动映射为唯一值(10 的幂次)

Table 4 Activity is mapped to a unique value (power of 10)

| Activity | Value  |
|----------|--------|
| A        | $10^7$ |
| B        | $10^6$ |
| C        | $10^5$ |
| D        | $10^4$ |
| E        | $10^3$ |
| F        | $10^2$ |
| G        | $10^1$ |
| H        | $10^0$ |

利用下式可以为每一个片段  $l\text{-}gram$  分配独一无二的数值表示:

$$value(l\text{-}gram) = \omega \times A_I + \sum_{j \neq I, j \neq 0} A_j + \lambda \times A_O$$

对图 4 中的片段  $l\text{-}gram$  分别计算  $value$ ,如表 5 所列。

表 5 片段  $l$ -gram 的  $value$  值  
Table 5  $value$  of clip  $l$ -gram

| $l$ -gram | Mapping Calculation                           | $value$  | $l$ -gram | Mapping Calculation                           | $value$  |
|-----------|-----------------------------------------------|----------|-----------|-----------------------------------------------|----------|
| HEAB      | $2 \times 10^0 + 10^3 + 10^7 + 3 \times 10^6$ | 13001002 | EABC      | $2 \times 10^3 + 10^7 + 10^6 + 3 \times 10^5$ | 11302000 |
| EABD      | $2 \times 10^3 + 10^7 + 10^6 + 3 \times 10^4$ | 11032000 | ABCD      | $2 \times 10^7 + 10^6 + 10^5 + 3 \times 10^4$ | 21130000 |
| ABDG      | $2 \times 10^7 + 10^6 + 10^4 + 3 \times 10^1$ | 20010030 | BCDG      | $2 \times 10^6 + 10^5 + 10^4 + 3 \times 10^1$ | 21100300 |
| BDGF      | $2 \times 10^6 + 10^4 + 10^1 + 3 \times 10^2$ | 2010310  | CDGF      | $2 \times 10^5 + 10^4 + 10^1 + 3 \times 10^2$ | 210310   |
| HABC      | $2 \times 10^0 + 10^7 + 10^6 + 3 \times 10^5$ | 11300002 | EHAC      | $2 \times 10^3 + 10^0 + 10^7 + 3 \times 10^5$ | 10302001 |
| ABCD      | $2 \times 10^7 + 10^6 + 10^5 + 3 \times 10^4$ | 21130000 | HACB      | $2 \times 10^0 + 10^7 + 10^5 + 3 \times 10^6$ | 10310002 |
| BCDF      | $2 \times 10^6 + 10^5 + 10^4 + 3 \times 10^2$ | 2110300  | ACBD      | $2 \times 10^7 + 10^5 + 10^6 + 3 \times 10^4$ | 21130000 |
| CDFG      | $2 \times 10^5 + 10^4 + 10^2 + 3 \times 10^1$ | 210130   | CBDF      | $2 \times 10^5 + 10^6 + 10^4 + 3 \times 10^2$ | 1210300  |
| HEAB      | $2 \times 10^0 + 10^3 + 10^7 + 3 \times 10^6$ | 13001002 | BDFG      | $2 \times 10^6 + 10^4 + 10^2 + 3 \times 10^1$ | 2010130  |

从表 5 可以得到事件日志  $L_{ch}$  中的高频片段组合为  $AB$ - $CD$ ,即理论上认为系统执行活动  $A, B, C, D$  时所需要的成本要高于执行其他活动所花费的成本,因此在迹对齐时此类活动的成本函数值要高于其他的活动。

接下来根据算法 2 给出具体分析。考虑第 2 节中的源日志  $L_{or}$ ,首先对其进行预处理,把发生频率极低的迹过滤掉,把相同发生轨迹的迹合并到一起,从而得到处理后的源日志  $L_{or}'$ ,如表 6 所列。

表 6 预处理后的源日志  $L_{or}'$   
Table 6 Preprocessed source log  $L_{or}'$

| 实例 | 迹 $\sigma_j (1 \leq j \leq 5)$ | 实例数 |
|----|--------------------------------|-----|
| 1  | HABDGF                         | 190 |
| 2  | HABCDGF                        | 219 |
| 3  | HABDFG                         | 288 |
| 4  | HACBDFG                        | 398 |
| 5  | HABCDGF                        | 523 |

第 4 节已经计算出  $L_{ch}$  中的高频片段  $l$ -gram 为  $ABCD$ 。基于此,对函数  $S(a, b), I(a, b)$  进行赋值。对活动  $A, B, C, D$  而言,令  $S(a, b) = 2(a = b)$ ,否则  $S(a, b) = -2$ ,且  $I(a, b) = -2$ ;日志  $L_{ch}$  中的其他活动则有相对更低的成本,即  $S(a, b) = 1(a = b)$ ,否则  $S(a, b) = -1$ ,且  $I(a, b) = -1$ 。

接下来计算  $L_{ch}'$  中的迹  $t_i$  与源日志  $L_{or}'$  的相似性度。以迹  $t_1$  为例,首先分别计算  $t_1$  与  $\sigma_j (1 \leq j \leq 5)$  的最优对齐,  $t_1$  与  $\sigma_1$  的最优对齐如表 7 所列。

表 7  $t_1$  与  $\sigma_1$  的最优对齐  
Table 7 Optimal alignment with  $t_1$  and  $\sigma_1$

| $t_1$ | $\sigma_1$ |
|-------|------------|
| H     | H          |
| E     | —          |
| A     | A          |
| B     | B          |
| D     | D          |
| G     | G          |
| F     | F          |

由此计算得出  $Score(t_1, \sigma_1) = 8$ ;同理依次计算得出  $Score(t_1, \sigma_2) = 1, Score(t_1, \sigma_3) = 5, Score(t_1, \sigma_4) = 3, Score(t_1, \sigma_5) = 6$ 。根据算法 2 计算出  $f_{cost}(t_1, \sigma_j) = \sum_{j=1}^5 Score(t_1, \sigma_j) = (8 + 1 + 5 + 3 + 6) \times 310 = 7130$ ,即  $f_{cost}(t_1, L_{or}) = 7130$ ;进而计算出  $t_1$  与  $L_{or}$  的相似性度为:

$$Sim(t_1, L_{or}) = \frac{f_{cost}(t_1, L_{or})}{|t_1| \times (|\sigma_1| + |\sigma_2| + |\sigma_3| + |\sigma_4| + |\sigma_5|)}$$

$$= \frac{7130}{7 \times (190 + 219 + 288 + 398 + 523)}$$
$$= 0.6295$$

根据上述计算步骤,直接展示  $L_{ch}$  中的剩余迹与  $L_{or}$  的计算结果。表 8 列出了  $t_2, t_3, t_4$  与  $\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5$  对齐的  $Score$  值。

表 8  $t_i$  与  $\sigma_j$  对齐的  $Score$  值  
Table 8 Aligned  $Score$  value of  $t_i$  and  $\sigma_j$

|       | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | $\sigma_5$ |
|-------|------------|------------|------------|------------|------------|
| $t_2$ | 4          | 11         | 7          | 5          | 8          |
| $t_3$ | 6          | 7          | 3          | 1          | 10         |
| $t_4$ | 3          | 4          | 6          | 10         | 1          |

利用表 8 的数据可以进一步计算  $t_2, t_3, t_4$  与  $L_{or}$  的总成本  $f_{cost}$  和相似性度  $Sim$ ,如表 9 所列。

表 9  $t_i$  与  $L_{or}$  的成本值和相似性度  
Table 9 Cost value and similarity degree of  $t_i$  and  $L_{or}$

|                 | 总成本 $f_{cost}$ | 相似性度 $Sim$ |
|-----------------|----------------|------------|
| $(t_1, L_{or})$ | 7130           | 0.6295     |
| $(t_2, L_{or})$ | 10115          | 0.8931     |
| $(t_3, L_{or})$ | 12339          | 0.9533     |
| $(t_4, L_{or})$ | 12096          | 0.9345     |

根据算法 2 和表 9 中的数据可以得出  $L_{ch}$  中与  $L_{or}$  相似性度比较低的是迹  $t_1$ 。理论上认为,  $Sim$  为 1 时相似性度最高,也可以认为完全一致,但是在系统实际运行中由于某些客观原因,相似性度大多无法达到 1,因此本文把相似性度阈值设为  $0.8(\beta = 0.8)$ ,只要  $Sim > \beta$ ,就认为相似性度是比较高的,即系统前后没有发生变化。对表 9 中的相似性度进行排序:  $Sim(t_1, L_{or}) < Sim(t_2, L_{or}) < Sim(t_4, L_{or}) < Sim(t_3, L_{or})$ ,并且  $Sim(t_1, L_{or}) < \beta$ ,因此可得到与源日志  $L_{or}$  相比  $L_{ch}$  中发生变化的迹为  $t_1$ 。到此为止,根据算法 1 和算法 2 挖掘出了系统变化前后发生变化的迹。

分析以上实例可以看出,与文献[13]的方法相比,本文中所采用的变化挖掘方法更具有完整性。

**结束语** 业务流程的变化对业务流程管理至关重要。本文在没有参考模型的情况下,基于事件日志的最优对齐对系统产生的变化给出精确定位分析。首先,提出一种从日志中挖掘高频形态学片段的算法,对变化后的日志进行挖掘,得到一组高频发生片段,因此对该片段赋予较高的成本函数值,进而找到变化日志中的迹与源日志的最优对齐。接着,基于最优对齐提出变化迹挖掘算法,并计算变化日志中的每一条迹与源日志的相似性度,通过比较得出相似性度较低的迹为发

生变化的迹,最终挖掘出变化迹集,并用实例显示了该算法的实用性。实例分析表明了本文算法挖掘出的变化的完整性,其在流程模型优化时起到了不可替代的作用。

然而,本文算法仍存在一些不足,如没有考虑当源日志和变化日志均为不完备日志时,如何精确定位出发生变化的完整迹集。因此,我们未来将主要致力于不完备日志下的迹变化挖掘。

## 参 考 文 献

- [1] WEIDLICH M, WESKE M, MENDLING J. Change propagation in process models using behavioural profiles[C]//2009 IEEE International Conference on Services Computing. IEEE, 2009: 33-40.
- [2] HONG Y K, ZHAO W, YAO Q. Model of BPMS for supporting variability of business process based on dependence[J]. Application Research of Computers, 2009, 26(9): 3348-3353, 3357.
- [3] FANG X W, ZHAO F, FANG H, et al. The fusion analysis method about the change region of the business process model based on behavior inclusion in petri net[J]. Chinese Journal of Computers, 2018, 41(3): 695-708.
- [4] GROSSMANN G, MAFAZI S, MAYER W, et al. Change propagation and conflict resolution for the Co-evolution of business processes[J]. International Journal of Cooperative Information Systems, 2015, 24(1): 1540002.
- [5] FDHILA W, RINDERLE-MA S, REICHERT M. Change propagation in collaborative processes scenarios[C]//Proceedings of the 8th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing. IEEE, 2012: 452-461.
- [6] ZHANG H, LI J, ZHU L M, et al. Investigating dependencies in software requirements for change propagation analysis[J]. Information and Software Technology, 2014, 56(1): 40-53.
- [7] FANG X W, PENG K, WANG L L, et al. Analysis of business process change propagation based on configuration and cancellation state[J]. Computer Integrated Manufacturing Systems, 2018, 24(7): 1621-1630.
- [8] DAHMAN K, CHAROY F, GODART C. Alignment and change propagation between business processes and service-oriented architectures[C]//2013 IEEE International Conference on Services Computing. IEEE, 2013: 168-175.
- [9] DAM H K, LE L S, GHOSE A. Supporting change propagation in the evolution of enterprise architectures[C]//2010 14th IEEE International Enterprise Distributed Object Computing Conference. IEEE, 2010: 24-33.
- [10] FANG H, ZHANG Y, WU Q L. A log induced change mining method for fault diagnosis using structure causality in BPMSs[J]. Control Theory and Technology, 2018, 35(8): 1167-1176.
- [11] GÜNTHER C W, RINDERLE S, REICHERT M, et al. Change mining in adaptive process management systems[M]//On the Move to Meaningful Internet Systems 2006: CoopIS, DOA, GADA, and ODBASE. Berlin: Springer, 2006: 309-326.
- [12] LI C, REICHERT M, WOMBACHER A. Mining business process variants: Challenges, scenarios, algorithms[J]. Data & Knowledge Engineering, 2011, 70(5): 409-434.
- [13] LI C, REICHERT M, WOMBACHER A. Discovering reference models by mining process variants using a heuristic approach[M]//Lecture Notes in Computer Science. Berlin: Springer, 2009: 344-362.
- [14] SONG M, GÜNTHER C W, VAN DER AALST W M P. Trace clustering in process mining[M]//Business Process Management Workshops. Berlin: Springer, 2009: 109-120.
- [15] HALLERBACH A, BAUER T, REICHERT M. Configuration and management of process variants[M]//Handbook on Business Process Management 1. Berlin: Springer, 2010: 237-255.
- [16] FANG H, SUN S Y, FANG X W. Behavior change mining methods based on incomplete logs conjoint occurrence relation[J]. Computer Integrated Manufacturing Systems.
- [17] JAGADEESH C B R P, VAN DER AALST W M P. Process diagnostics using trace alignment: Opportunities, issues, and challenges[J]. Information Systems, 2012, 37(2): 117-141.
- [18] HUANG D M, XU Q Q, HE Q, et al. Workflow model based on TC-Net[J]. Computer Engineering and Applications, 2016, 52(15): 260-265.
- [19] VAN DE AALST W. Process discovery: capturing the invisible[J]. IEEE Computational Intelligence Magazine, 2010, 5(1): 28-41.
- [20] POURMASOUMI A, KAHANI M, BAGHERI E. Mining variable fragments from process event logs[J]. Information Systems Frontiers, 2017, 19(6): 1423-1443.



**LIU Jing**, born in 1995, postgraduate. Her main research interests include Petri net and change mining.



**FANG Xian-wen**, born in 1975, Ph.D., professor. His main research interests include Petri net and trusted software.