Journal of System Simulation

Volume 34 | Issue 7 Article 16

7-20-2022

# Interactive Construction of Scientific Workflow Based on Process Mining

Jun Liu
*1.Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China;2.University of Chinese Academy of Sciences, Beijing 100049, China;*, liujun@sccas.cn

Yang Gao
*1.Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China;*

Tao Xu
*1.Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China;*

Qing Zhao
*1.Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China;*

*See next page for additional authors*

# Interactive Construction of Scientific Workflow Based on Process Mining


## Abstract

Abstract: When dealing with large-scale or complex workflows, the construction efficiency of traditional interactive workflow construction methods is very low. *To solve this problem, a workflow construction method based on process mining is proposed. Heuristic methods are used to collect process fragments. The specially designed relation description language is used to record the process description of different levels and aspects in the workflow as text. The text is translated to generate process relational data, which will be output to the process discovery algorithm to generate a sound workflow network.* An interactive workflow construction software has been developed and tested in scientific workflow cases to verify the effectiveness of this method and show better interactive efficiency compared with other methods.



## Keywords

scientific workflow, workflow modeling, process mining, relation description language, human-computer interaction



## Authors

Jun Liu, Yang Gao, Tao Xu, Qing Zhao, Guihua Shan, and Xuebin Chi



## Recommended Citation

Jun Liu, Yang Gao, Tao Xu, Qing Zhao, Guihua Shan, Xuebin Chi. Interactive Construction of Scientific Workflow Based on Process Mining[J]. Journal of System Simulation, 2022, 34(7): 1547-1558.

第 34 卷第 7 期 系统仿真学报© Vol. 34 No. 7
2022 年 7 月 Journal of System Simulation Jul. 2022

# 基于过程挖掘的科学工作流交互式构建

刘俊[1,2]，高阳[1]，徐涛[1]，赵青[1]，单桂华[1]，迟学斌[1,2]
(1. 中国科学院 计算机网络信息中心，北京 100190；2. 中国科学院大学，北京 100049)

**摘要：**传统的交互式工作流构建方法在处理规模较大或流程复杂的工作流时构建效率较低。为此，*提出了一种基于过程挖掘的工作流构建方法，采用启发式问答方式收集记录流程片段，通过特别设计的关系描述语言以实现工作流中不同层级和不同环节的过程描述，由描述文本转译生成的流程关系数据被输出到过程发现算法中，从而生成健全的工作流网络。*实现了一套交互式工作流构建工具，并在科学工作流案例中进行了测试，验证了方法的有效性，在与其它方法相对比时表现出更佳的交互效率。

**关键词：**科学工作流；工作流构建；过程挖掘；关系描述语言；人机交互
中图分类号：TP391 文献标志码：A 文章编号：1004-731X(2022)07-1547-12
DOI: 10.16182/j.issn1004731x.joss.21-0151

## Interactive Construction of Scientific Workflow Based on Process Mining

*Liu Jun*[1,2], *Gao Yang*[1], *Xu Tao*[1], *Zhao Qing*[1], *Shan Guihua*[1], *Chi Xuebin*[1,2]
(1. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China)

**Abstract:** When dealing with large-scale or complex workflows, the construction efficiency of traditional interactive workflow construction methods is very low. *To solve this problem, a workflow construction method based on process mining is proposed. Heuristic methods are used to collect process fragments. The specially designed relation description language is used to record the process description of different levels and aspects in the workflow as text. The text is translated to generate process relational data, which will be output to the process discovery algorithm to generate a sound workflow network.* An interactive workflow construction software has been developed and tested in scientific workflow cases to verify the effectiveness of this method and show better interactive efficiency compared with other methods.

**Keywords:** scientific workflow; workflow modeling; process mining; relation description language; human-computer interaction

## 引言

工作流是指由一系列可重复的活动有机联系组成的程序模式，可在资源的支撑驱动下实现材料的转换、服务的提供以及信息的处理。工作流的应用涉及众多领域，如科学计算、工业自动化、互联网以及物联网等。在科学计算领域存在很多工作流管理工具，如 Taverna[1]，Kepler[2]，Galaxy[3]等，可提供科学工作流的设计、构建、执行、存档、协作等功能。除此之外，还可通过 RapidMiner[4]和 KNIME[5]等工具，针对科学家需求进行量身定制，以实现工作流在集成环境下的数

收稿日期：2021-03-01 修回日期：2021-04-22
基金项目：国家重点研发计划（2017YFB0203504）
第一作者：刘俊（1982-），男，硕士，高工，研究方向为可视化及人机交互。E-mail：liujun@sccas.cn

据挖掘、预测分析、业务优化、实施报告等功能。借助适当的工具，工作流技术的应用可极大提高科学计算的自动化水平及服务能力。

在未采用工作流框架的情况下，很多科学计算应用在发展到一定规模后，由于系统复杂度的不断增加，总体性能达到了瓶颈，计算应用的发展速度无法得到有效保证，要实现应用服务的自动化和智能化难上加难。为了有效保证科学计算的实施效率，通过流程细化及自动化，提升计算应用的智能化水平，在原有的计算应用基础上进行工作流建模，构建可兼容科学工作流管理体系的显式工作流模型非常必要。但是，构建工作流是一个非常复杂且耗时费力的过程，难以找到一个完全了解系统中全部过程的人员，因为参与系统的每个人通常只熟悉整个系统的部分领域，而让具有不同技能和背景的人员共同参与工作流管理又将带来复杂的沟通问题。较为高效的方法是指派特定的工作流管理人员，与系统中各领域专业人员面谈并录入流程片段，同时采用交互式的工作流构建工具，在录入过程中快速反馈模型构建结果，以实现结果确认[6]。因此，我们提出了一种基于过程挖掘的工作流构建方法。该方法通过启发式问答收集记录流程片段，采用关系描述语言实现工作流中不同层级和不同环节的过程描述，通过语言解析生成关系矩阵数据，并采用过程发现算法生成健全的工作流网络。

# 1 相关工作

## 1.1 工作流模型构建

过程建模对于科学工作流管理乃至商业过程管理来说至关重要，而工作流模型的设计和统一是高效流程分析管理的重要开端。在工作流模型理论基础上，运营管理及运筹学等管理科学得以建立，流程的推理设计以及基于流程的决策、计划和控制等也得以开展。在过程挖掘学科范畴内，过程是指一系列可执行动作的组合[7]，而工作流模型则是对动作组合成为过程的方法描述。一般情况下，工作流模型可以用有向图的形式进行表达，其中图的节点用以表示工作流的状态或者动作，图的边则表示动作之间的潜在控制流，例如Petri网[8]，WF网[9]，BPMN(business process model and notation)[10]、Casual网[11]都是常用的工作流模型。

大部分交互式的工作流构建工具均采用拖拽式的流程构建方法，如Taverna、Kepler、AProMore提供了强大的图形化建模工具，可通过图形界面排布一系列的工作流模块，让用户自由选择及联接组合。这种构建方法较为简单直接，但对于规模较大或流程复杂的工作流来说，构建效率较低，需要适当的交互硬件条件支持。在没有模型专业知识的情况下，这种方法无法保证输出模型的健全性及精确性。而如果让专业的流程管理人员进行辅助建模，则真正了解具体流程的领域专家只能作为建模过程的旁观者，无法在构建过程中与模型实现互动[12-13]。因此，本文采用易于理解的图形界面，同时简化了建模过程，加大了领域专家的介入程度，以提高建模效率。

基于文本的工作流构建方法也是较常用的方法之一。如Graphviz[14]采用简单的顺序关系建模，VipTool[15]和Oclets[16]采用事例片段记录来创建工作流模型。这类方法由于缺乏对建模过程的控制，其建模结果容易受到记录中的噪声及错误的影响，从而产生大的偏差。为了减少噪声的干扰，一系列结构化的建模方法被提出。如TAD(tabular application development)[13]让业务分析人员从活动表开始建模，然后再转换为图形化的流程图控制。Sketch Miner[6]采用规范化的领域专用语言DSL(domain specific language)来描述事例以建立模型。然而基于结构化描述的建模方法学习成本高，且输入数据与工作流模型紧切相关，灵活性较差。为了将工作流可视化界面与事例记录分析数据相结合，Little Thumb[17]基于可视化界面结合启发式的建模过程进行过程发现。IPD(interactive process discovery)[18]采用交互式的过程发现方法，利用专




DOI: 10.16182/j.issn1004731x.joss.21-0151

家知识去除输入数据中的噪声和错误，保证工作流结果模型的结构性及健全性。但其目标是对庞大的日志数据进行统计，发现已有工作流而非建立新的工作流，且其可提供的交互方式非常有限。

### 1.2 过程挖掘

过程挖掘[19]是一类工作流模型[7]发现方法，旨在通过大量过程运行日志训练得到该过程的工作流模型。其中过程运行日志指的是运行过程中触发的事件所组成的序列。由过程挖掘方法得到的结果模型应尽量符合日志中出现的情况，即应尽量覆盖全部日志事件序列(适用性)，同时不产生超出日志外的更多的事件序列(精确性)。另外，计算产生的结果模型应尽量简化(概括性)，但过度简化的模型通常在精确性上不够理想[7,19-21]。由于日志是从实际流程执行中收集的，因此该训练过程只能在正向样本中进行[7]，但对于日志中可能存在的噪声(异常事件和低频事件)，通常的处理方法是尽量忽略这类数据[19]。

在过程挖掘学科领域发展早期，基于WF网模型，文献[9,19]中给出的过程发现算法取得了里程碑式的成功，其后发展出的α算法家族[22-23]是领域内知名度最高的算法。该类算法计算快速且稳定性高，其挖掘出的结果模型被证明在健全度及结构化程度方面均有所保证。

启发式算法[24-26]是另一类较有代表性的过程发现算法，这类算法可有效降低日志中存在的噪声对挖掘结果的干扰，而且通常可挖掘出相比于其它算法结果结构化程度更优的模型。基于线性规划的ILP(integer linear programmin)算法[27]得到的结果在适用性上表现较好，但算法稳定性较差，且平均计算时间依赖于其所使用的变量数量，因此不适用于情况较复杂的输入日志处理。过程树[28]是一种专用于过程发现的结构化过程模型，可保证结果模型的健全性。归纳算法[29]和进化树算法[30-31]都是基于过程树模型的过程发现算法，其中归纳法通过挖掘模型中的切割关系来构建过程树，而进化树算法采用遗传算法，根据用户设定的模型质量维度偏好来驱动发现过程。进化树算法的计算时间相对较长，但其结果模型在精度和适用性方面均有较强的竞争力。

由上可见，相比于其它过程发现算法，α算法家族所采用的WF网工作流模型更为通用，便于检查结果模型的健全性以及质量评估，结果模型通过简单变换处理可转换为BPMN等语法较复杂的网络。在计算效率、精确性、稳定性方面α算法家族表现出众，在适用性、概括性等方面也表现出一定优势，较为适合在要求快速反馈且计算资源水平较低的交互场景中使用。因此在本文中，我们将基于α算法思想来实现基于活动关系矩阵的过程发现。

## 2 背景说明

本节将介绍WF网工作流模型相关的一些概念。因为篇幅限制，假设读者已对Petri网相关概念有一定了解。文中采用$N=(P,T,I,O)$表示Petri网，其中$P$为网络中的位置集合，$T$为网络中的迁移活动集合，$I$为输入函数，$O$为输出函数。采用$\bullet p=\{t|O(t,p)=1\}$为位置$p$的前续迁移活动集合，采用$p\bullet=\{t|I(p,t)=1\}$为位置$p$的后继迁移活动集合。

定义1：WF网[9]是一类特殊的Petri网：① 它具有2个特殊位置，即源点位置$i$以及汇点位置$o$，其中$\bullet i=\varnothing$，且$o\bullet=\varnothing$；② 如果这个网络中添加一个迁移活动以建立位置$o$到$i$的连接，将得到一个强连通图。

WF网具有与Petri网相同的标记触发机制，若在标记$M$状态下，则活动$t$可被触发，记作$M\geqslant t$。

定义2：一个WF网$N=(P,T,I,O)$是健全的，当且仅当在初始标记$M_0=[i]$下，$(N,M_0)$符合以下条件：① 该网络是安全的，即对任意$p\in P$，在经过若干次触发后得到的标记满足$M_i(p)\leqslant 1$；



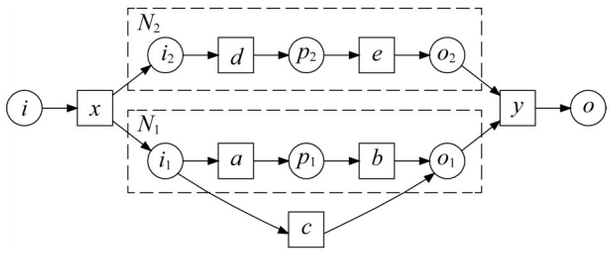
② 对于任意属于$M_0$可达集$R(N,M_0)$的标记$M_i$，当$o \in M_i$时则有$\boldsymbol{M}_i = [o]$；③ $N$中不存在无活的迁移活动，即对于任意迁移活动$t \in T$，存在一个标记$M_k \in R(N,M_0)$使得$M_k \geqslant t$。

如图1(a)所示的健全WF网$N$，连续的活动触发序列$\sigma = \langle x,d,e,a,b,y \rangle$可使系统从初始标记状态$\boldsymbol{M}_0 = [i]$进入最终标记状态$\boldsymbol{M}_e = [o]$，这样的序列被称为工作流事例，记作$\sigma \to N$。对于$T^* \subset T$，以及事例$\sigma$，设$|\sigma| = n$，将满足条件$\forall_{k \leqslant i \leqslant m}\sigma(i) \in T^*$，且 $k \neq m \wedge m = n \vee \sigma(m+1) \notin T^* \wedge k = 1 \vee \sigma(k-1) \notin T^*$的$k$和$m$找出，将$\sigma$中由第$k$个至第$m$个元素组成的序列替换为活动标签$t$，得到一个新的活动触发序列$\sigma'$，结果记为$\sigma' = \sigma^{[T^* \backslash t]}$。

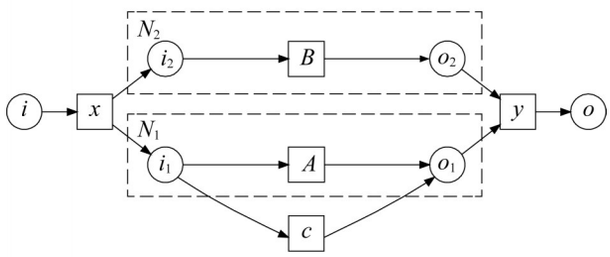
定义3：对于WF网$N = (P,T,I,O)$中的迁移活动$a,b \in T$：当且仅当存在标记$M_k \in R(N,M_0)$且$M_k \geqslant a$在触发活动$a$后到达的标记状态$M_{k+1} \geqslant b$，则$a$与$b$的关系被称为跟随关系，记作$a \to b$。当且仅当存在位置$p \in P$满足$O(a,p) = 1$且$I(p,b) = 1$，则$a$与$b$的关系被称为顺序关系，记作$a \Rightarrow b$。当且仅当存在位置$p_1,p_2 \in P$满足$I(p_1,a) = 1$，$I(p_1,b) = 1$，$O(a,p_2) = 1$且$O(b,p_2) = 1$，则$a$与$b$的关系被称为选择关系，记作$a \vee b$。当且仅当存在迁移活动$t_1,t_2 \in T$满足$t_1 \Rightarrow a$，$t_1 \Rightarrow b$，$a \Rightarrow t_2$且$b \Rightarrow t_2$，则$a$与$b$的关系被称为并发关系，记作$a \wedge b$。

定义4：对于WF网$N = (P,T,I,O)$，存在$N' = (P',T',I',O')$以及$\varnothing \neq T^* \subseteq T$，$t^* \in T'$，满足$T^* = T \backslash T'$，$\{t^*\} = T' \backslash T$，且$\forall_{\sigma \to N}\sigma^{[T^* \backslash t^*]} \to N'$，则称活动$t^*$是$t \in T'$的超级活动，记作$t^* \supset t$。

如图1(a)所示的WF网$N$，$N_1$和$N_2$均为其子网，经过源点和汇点的重新连接，$N_1$和$N_2$成为如图1(b)所示的只有一个迁移活动的WF网，此时，$A$与$a$为包含关系($A \supset a$)，$B$与$d$为包含关系($B \supset d$)，$x$和$A$为顺序关系($x \Rightarrow A$)，$A$和$B$为并发关系($A \wedge B$)，$A$和$c$为选择关系($A \vee c$)。




(a) 健全WF网




(b) 利用超级活动简化后的WF网

图1 健全WF网及其子网的迁移活动关系示例

Fig. 1 Example for the transition relations of a sound WF network and its subnets

## 3 方法介绍

### 3.1 设计思路

采用过程发现算法通过事例集生成健全的工作流需要大量的、完整的且准确的事例集轨迹数据，而在以手工输入为数据源的条件下，采集到的数据往往具有相反的特征。为了提高手工输入数据的效率，本文提出一套方法，将手工输入的数据直接转换成过程发现算法中的中间数据，跳过事例集轨迹分析过程，避免了对冗余信息的不必要的分析处理，并加入自动化的流程关系挖掘步骤，从用户输入数据中找到潜藏的流程关系。

如图2所示，首先用户通过手工录入关系描述语句的方式为过程发现算法提供活动集合，由关系描述文本转译生成模型发现所需的活动关系矩阵。由于用户输入数据的不完整，存在未被记录的活动间流程关系，这种情况下得到的活动关系矩阵会导致挖掘结果模型不健全。此时，采用本文提出的流程关系挖掘以及超级活动展开算法，可找到活动之间潜藏的并发关系以及顺序关系，通过处理生成符合要求的关系矩阵。为了弥补手




DOI: 10.16182/j.issn1004731x.joss.21-0151

工输入可能出现的人为错误，用户可通过交互界面对关系描述文本进行即时调整，经过实时的模型结果生成及反馈，可实现建模过程的交互。

经过上述方法流程后，用户将得到基于WF网理想模型的工作流，此时用户可在模型结果基础上进行模型扩展加工，如将模型变换为BPMN模型或与工作流引擎兼容的其它模型，或将流图的节点转换为任务或事件并使用角色注释将任务和事件放置在各自的泳道中以实现角色分工，或进行整体分层布局提高模型的可理解性等。以上操作与具体的工作流应用的运行环境及场景有紧密关系，而实现模型生成是模型应用的重要基础。

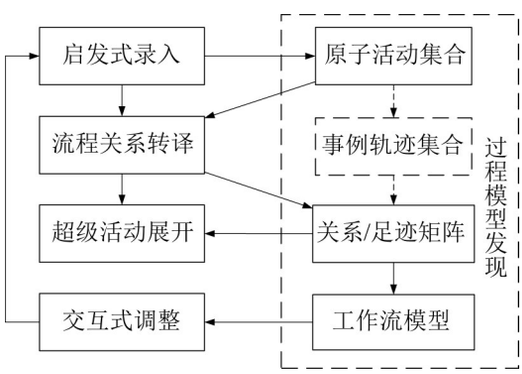



图2 基于过程挖掘的工作流构建方法流程图

Fig. 2 Flow chart of workflow construction method based on process mining

## 3.2 启发式录入

复杂的科学计算过程通常涉及诸多阶段或环节类型，尤其是对于分工日益细化的工作流应用场景，完成整个过程的工作流建模往往需要结合多方面的专家知识。为了保证建模效率同时提高模型的可维护性，有必要让整体过程中具体活动的相关主体，即特定的领域专业人员加入到工作流的协同建模过程中来。具体的操作方式可以是由工作流管理人员与领域专业人员进行面谈记录，或者由领域专业人员直接作为模型操作人员自行进行流程录入。采用特定语法的关系描述语言RDL(relation description language)实现流程的记录，其语法采用EBNF语言描述如下：

<关系描述文本> ::= (<关系描述语句> 'EOL')$^+$ 'EOF'

<关系描述语句> ::= <跟随语句>|<选择语句>|<断开语句>

<跟随语句> ::= [[<活动标签>]':']<活动标签>((',';')<活动标签>)*

<选择语句> ::= <活动标签>('|'<活动标签>)$^+$

<断开语句> ::= <活动标签>('#'<活动标签>)$^+$

<活动标签> ::= (一般字符或空格)$^+$

跟随语句$f$为形如$a{:}t_1r_1t_2r_2\cdots r_{n-1}t_n(t_i \in T,\ r_i$为逗号或分号)的字符串，按是否包含冒号形式的分解符分为分解型和片段型两类。

分解型的跟随语句在记录活动$t_{i-1}$与$t_i(2 \leqslant i \leqslant n)$的顺序关系以及并发关系的同时，采用冒号形式的分解符，以表明活动序列$t_1,t_2,\cdots,t_n$是活动$a$的分解，活动$a$是语句标出的所有活动$t_i(1 \leqslant i \leqslant n)$的超级活动，即$a \supset t_i$。对于该语句来说，$a$为$f$所描述的超级活动，记作$\text{super}(f)=a$，活动序列的首末活动代表了该语句所记录的事例的入口活动和出口活动，记作$\text{first}(f)=t_1$及$\text{last}(f)=t_n$。当跟随语句中超级活动未被标出时，记作$\text{super}(f)=\odot$，这表明该语句所描述的是工作流全局运行过程的一个事例，又称为全局事例。跟随语句中逗号分隔符表示前后相邻的活动$a$和$b$之间存在顺序关系，即$a \Rightarrow b$，而使用分号分隔时则表示前后相邻活动$a$和$b$之间存在并发关系，即$a \wedge b$。

片段型的跟随语句不采用冒号分解符，仅表达相邻活动间的顺序关系或并发关系，表示方法与上述相同。

选择语句被用于表示语句中全部活动之间的选择关系。断开语句则被用于表示语句中相邻活动不存在顺序关系，以消除在跟随语句中过度声明而记录的多余顺序关系。

本文采用启发式问答来完成不同层级和环节



的叙事建模记录。具体步骤为

(1) 安排对目标活动流程的整体情况有所掌握的人员进行系统问答；

(2) 询问流程中最常见的活动安排和执行顺序，记录活动进行过程中是否存在可并发执行的环节，并通过跟随语句进行描述，若该流程涉及多重关系，则采用超级活动标签标记，以备其它语句引用；

(3) 询问各活动是否有可替代的活动，并通过选择语句进行描述；

(4) 询问在前述活动过程描述中是否存在被忽略的禁制条件，并通过断开语句进行修正；

(5) 询问各活动是否是可分解的，如果可分解，则针对该活动采用本流程进行更细粒度的启发式问答记录。

需要注意的是，为保证建模结果的健全性，采集到的关系描述文本应具备以下几个条件：

(1) 关系描述文本中应存在描述全局事例的跟随语句；

(2) 全部跟随语句中出现的活动可按包含关系和选择关系追溯到某条描述全局事例的跟随语句；

(3) 全部片段型跟随语句的首个活动以及最后一个活动应出现于某条跟随语句中。

另外，由于参与人员的偏好以及个人视角的限制，单次采集的记录数据往往是不完整的，或存在偏向性，从而不能完全反映工作流的全貌，因此，除了需要对不同层级和环节的领域专业人员进行记录采集，还需要对同样层级和环节但不同偏好的人员进行采集。

### 3.3 流程关系转译

经过上述的建模记录过程，可获得以RDL记录的流程片段和活动间流程关系的描述。为了产生过程模型发现算法可接受的输入数据，后续将对关系描述文本所记录的流程关系进行转译，以将其转换为工作流模型发现算法所需的流程关系矩阵。

按分解型跟随语句的定义，语句中标出的超级活动可按此语句主体部分展开成相应的活动序列。而该活动序列中的活动也可能在其它跟随语句的声明支撑下被进一步展开。在得到完整的关系描述文本时，可以确定哪些活动是不可展开的，这些活动被称为原子活动。

设$T$为关系描述文本中出现的所有活动的集合，$S_{\subset}$是文本所描述的包含关系的集合，则原子活动集合为$T_{\varepsilon}=\{a|\nexists|_{x\in T}(x,a)\in S_{\subset}\}$，相应的超级活动集合为$T_{\sigma}=T\backslash T_{\varepsilon}$。由于将关系描述文本进行转译得到的流程关系矩阵应只与原子活动有关，因此，为防止展开过程中出现循环死锁的情况，活动间的包含关系派生的有向图$(T,S_{\subset})$必须是无环的。

基于手工输入进行工作流建模常常存在数据准确性及完备性方面的问题(其中准确性要求指的是用户描述的事件关系应全部在生成的工作流中得到正确的体现，完备性要求指的是可由用户描述推理得到的潜藏事件关系也应在生成的工作流中得到完全的体现)。因此，为保证数据的准确性，应在对关系描述文本进行流程关系转译时，通过解析语法检查以及若干必要条件的判断，避免用户由于不完整输入或误输入造成的错误，而实时交互系统的图形反馈也可以让用户及时发现输入数据的异常点。

在完备性方面，用户出于主观或客观原因很难给出足够充分的关系描述，需要由已知数据进一步挖掘潜藏的流程关系。如图3所示，由用户输入的关系描述语句中，可以得到$b\vee c$以及$a\Rightarrow b$，由此可推出$a\Rightarrow c$，同理，由$b\vee c$以及$b\wedge d$可推出$c\wedge d$。总之，通过某些条件组合以相应的推理可挖掘出潜藏的一些流程关系，在本文所解决的问题中，较实用的推理条件如表1所示。

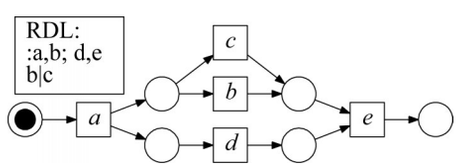



图3 流程关系挖掘示例

Fig. 3 Example for process relationship mining




DOI: 10.16182/j.issn1004731x.joss.21-0151

表1 流程关系挖掘推理条件列表
Table 1 List of inference conditions for process relation mining

| 条件 | 结果 |
|---|---|
| $a \vee b$且$a \Rightarrow c$ | $b \Rightarrow c$ |
| $a \vee b$且$c \Rightarrow a$ | $c \Rightarrow b$ |
| $a \vee b$且$a \wedge c$ | $b \wedge c$ |
| $a \vee b$且$a \vee c$ | $b \vee c$ |
| 在同一条跟随语句中$a \wedge b$且$a \Rightarrow c$ | $b \Rightarrow c$ |
| 在同一条跟随语句中$a \wedge b$且$c \Rightarrow a$ | $c \Rightarrow b$ |

综合以上分析，本文设计了一套将关系描述文本转译为流程关系矩阵的算法。该算法流程如算法1所示。

算法1：流程关系转译

(1) 过滤关系描述文本得到全部的活动标签，形成活动集合$T$，解析得到全部的分解型跟随语句$F$以及其它功能语句，同时得到全部描述全局事例的跟随语句$F_{\odot}=\{f|\text{super}(f)=\odot, f\in F\}$；

(2) 检查$F_{\odot}$是否为空，如果是则报出异常并退出；

(3) 将关系描述语句中相邻活动间的关系按标签间符号进行判断并记录，得到顺序关系集合$S_{\Rightarrow}$、并发关系集合$S_{\wedge}$、选择关系集合$S_{\vee}$、包含关系集合$S_{\subset}$以及被断开语句特别指出的非顺序关系集合$S_{\not\Rightarrow}$；

(4) 检查有向图$(T,S_{\subset})$是否为无环的，如果不是则报出异常并退出；

(5) 按解析得到的活动间流程关系组成初始流程关系矩阵$\boldsymbol{R}$，矩阵元素按如下设置：

$$\boldsymbol{R}_{u,v}=\begin{cases}\Rightarrow, & (u,v)\in S_{\Rightarrow}\wedge(v,u)\notin S_{\Rightarrow}\wedge(u,v)\notin S_{\wedge}\wedge \\ & (u,v)\in S_{\vee}\wedge(u,v)\notin S_{\not\Rightarrow}\\ \Leftarrow, & (v,u)\in S_{\Rightarrow}\wedge(u,v)\notin S_{\Rightarrow}\wedge(v,u)\notin S_{\wedge}\wedge \\ & (v,u)\in S_{\vee}\wedge(v,u)\notin S_{\not\Rightarrow}\\ \wedge, & (u,v)\in S_{\wedge}\vee((u,v)\in S_{\Rightarrow}\wedge(u,v)\in S_{\Rightarrow})\\ \vee, & (u,v)\in S_{\vee}\wedge(u,v)\notin S_{\Rightarrow}\wedge(u,v)\notin S_{\Rightarrow}\wedge \\ & (u,v)\notin S_{\wedge}\\ 0, & \text{其它}\end{cases}$$

(6) 通过表1所示的推理条件挖掘潜藏的流程关系，并记录到流程关系矩阵$\boldsymbol{R}$中。

由于WF网不支持嵌套结构，转译后得到的流程关系矩阵中包含的超级活动无法加入到建模结果中，因此，有必要在后续处理中将全部的超级活动展开成原子活动，以满足模型发现算法的输入参数要求。

### 3.4 超级活动展开

定义5：对于有向无环图的$G(V,E)$中节点$a\in V$，设其父节点集合$\Gamma_G(a)=\{x|(x,a)\in E\}$，其子节点集合$\Delta_G(a)=\{x|(a,x)\in E\}$，则定义该节点的秩为

$$\text{rank}_G(a)=\begin{cases}0, & \Gamma_G(a)=\varnothing\\ \max\limits_{x\in\Gamma_G(a)}(\text{rank}_G(x))+1, & \Gamma_G(a)\neq\varnothing\end{cases}$$

为了将转译得到的全部包含关系转换为可填入关系矩阵的原子活动间流程关系，需要将全部与超级活动有关的流程关系展开成原子活动间的流程关系，具体方法如算法2所示。由于该算法将频繁查询过滤关系矩阵以找到符合条件的活动，为表述方便，对活动$a\in T$以及流程关系$r$，这里记$\boldsymbol{R}(r,a)=\{x|\boldsymbol{R}_{x,a}=r\}$，$\boldsymbol{R}(a,r)=\{x|\boldsymbol{R}_{a,x}=r\}$。

算法2：超级活动展开

(1) 计算有向无环图$(T,S_{\subset})$中节点的秩，并按秩大小对全部超级活动进行排序，得到序列$Q$；

(2) 从序列$Q$中取出排在最前位置的秩最小的超级活动$a$，并找出全部描述活动$a$事例的跟随语句$F_a=\{f|\text{super}(f)=a, f\in F\}$；

(3) 将关系矩阵中$\boldsymbol{R}_{u,v}$置为并发关系$\wedge$，当且仅当$u\in\Delta_{(T,S_{\subset})}(a)$，$v\in\boldsymbol{R}(\wedge,a)$；

(4) 将关系矩阵中$\boldsymbol{R}_{u,v}$置为顺序关系$\Rightarrow$，同时将$\boldsymbol{R}_{v,u}$置为逆序关系$\Leftarrow$，当且仅当$v\in\boldsymbol{R}(a,\Rightarrow)$，$u\in\bigcup_{f\in F_a}\boldsymbol{R}(\wedge,\text{last}(f))\cup\boldsymbol{R}(\vee,\text{last}(f))$且$(u,v)\notin S_{\not\Rightarrow}$；

(5) 将关系矩阵中$\boldsymbol{R}_{u,v}$置为顺序关系$\Rightarrow$，同时将$\boldsymbol{R}_{v,u}$置为逆序关系$\Leftarrow$，当且仅当$u\in\boldsymbol{R}(\Rightarrow,a)$，$v\in\bigcup_{f\in F_a}\boldsymbol{R}(\wedge,\text{first}(f))\cup\boldsymbol{R}(\vee,\text{first}(f))$且$(u,v)\notin S_{\not\Rightarrow}$；

(6) 检查序列$Q$是否为空，否则转到(2)继续展



开处理。

经过超级活动展开可得到较完整的流程关系矩阵，要在此矩阵基础上构建健全的WF网模型，还需要去除一些与结果模型无关的活动，如全部超级活动应从矩阵中去除，另外按照健全WF网的定义，通过检查关系图$(T_\varepsilon, \Rightarrow)$的连通情况，剔除不满足连通条件的活动节点，从而保证结果模型质量。

### 3.5 模型发现

本文采用文献[22]中所述的α算法的模型发现算法思想，但是原始方法采用的输入是日志数据，需对算法各步骤进行调整才能适用于以关系描述文本为输入的工作流构建应用场景。调整后的算法如算法3所示。

算法3：工作流模型发现

(1) 计算初始的入口活动集合 $T_I^0 = \bigcup_{f\in F_\odot} \boldsymbol{R}(\wedge,\text{first}(f)) \cup \boldsymbol{R}(\vee,\text{first}(f))$；

(2) 迭代计算 $T_I^{k+1} = T_I^k \cup (\bigcup_{a\in T_\sigma^k \cap T_\sigma} \bigcup_{f\in F_a} \boldsymbol{R}(\wedge,\text{first}(f)) \cup \boldsymbol{R}(\vee,\text{first}(f))) \backslash T_\sigma$，直至 $T_I^k \cap T_\sigma = \varnothing$，最终结果设为 $T_I$；

(3) 计算初始的出口活动集合 $T_O^0 = \bigcup_{f\in F_\odot} \boldsymbol{R}(\wedge,\text{last}(f)) \cup \boldsymbol{R}(\vee,\text{last}(f))$；

(4) 迭代计算 $T_O^{k+1} = T_O^k \cup (\bigcup_{a\in T_\sigma^k \cap T_\sigma} \bigcup_{f\in F_a} \boldsymbol{R}(\wedge,\text{last}(f)) \cup \boldsymbol{R}(\vee,\text{last}(f))) \backslash T_\sigma$，直至 $T_O^k \cap T_\sigma = \varnothing$，最终结果设为 $T_O$；

(5) 找到全部相互为选择关系的原子活动集合 $\mathbb{U}_\vee = \{ A | A \subseteq T_\varepsilon \wedge A \neq \varnothing \wedge \forall_{a_1,a_2\in A} a_1 \vee a_2 \}$；

(6) 找到所有可能的中间位置 $X = \{(A,B) | A \in \mathbb{U}_\vee \wedge B \in \mathbb{U}_\vee \wedge \forall_{a\in A} \forall_{b\in B} a \Rightarrow b\}$；

(7) 去除非最优的位置 $Y = \{(A,B) \in X | \forall_{(A',B')\in X} A \subseteq A' \wedge B \subseteq B' \Rightarrow (A,B) = (A',B')\}$；

(8) 加入两个特殊位置 $P = \{ p_{(A,B)} | (A,B) \in Y \} \cup \{ i,o \}$；

(9) 从位置到迁移活动的弧 $I = \{(p_{(A,B)},b) | (A,B) \in Y \wedge b \in B\} \cup \{(i,t) | t \in T_I\}$；

(10) 从迁移活动到位置的弧 $O = \{(a,p_{(A,B)}) | (A,B) \in Y \wedge a \in A\} \cup \{(t,o) | t \in T_O\}$；

(11) 得到结果模型WF网 $N = (P,T_\varepsilon,I,O)$。

## 4 案例测试

基于上述方法，本文在Windows操作系统环境下，采用Qt和Graphviz[14]等技术架构设计开发了交互式工作流构建软件RDLModel。其工作界面如图4所示。通过在输入框输入RDL关系描述语句，软件可实时反馈工作流建模结果。为了验证该方法的性能效率，基于典型工作流示例以及实际科学工作流应用进行了案例测试。

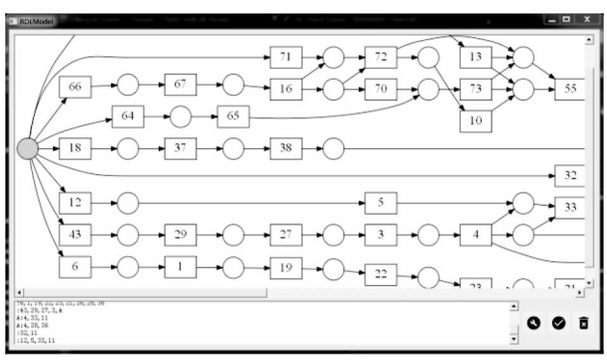



图4 RDLModel用户界面
Fig. 4 RDLModel user interface

### 4.1 典型工作流的构建

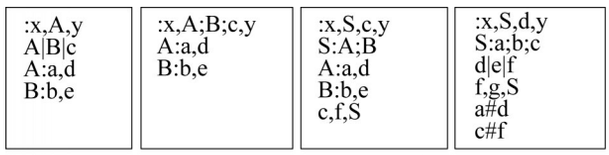
由于本文采用了α算法作为模型发现方法，因此，受到模型表达偏向的影响，只能生成可由WF网表达且被α算法支持的模型。图5选取了一些典型的WF网的构建案例，以说明不同类型的工作流的RDL描述方法。从图5中可看出，本文方法可很方便地支持多路径选择、多任务并发、循环跳转、混合结构工作流的构建。

在现有的工作中，与本文所采用的方法最接近的是Sketch Miner[6]所采用的方法。Sketch Miner方法采用自定义的DSL记录用户描述的工作流事例轨迹片段，其记录方式与关系描述文本中的跟随语句的语法类似，但是Sketch Miner不支持与选择语句以及断开语句类似的语法，同时Sketch

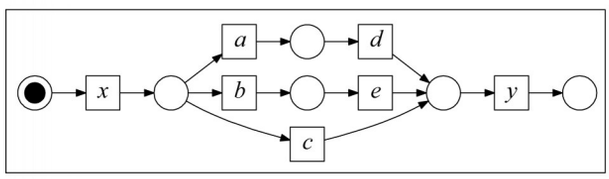



DOI: 10.16182/j.issn1004731x.joss.21-0151

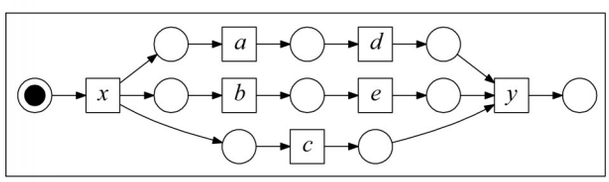
Miner虽然可通过省略符提供类似于超级活动描述方法，但灵活性和可读性较弱。



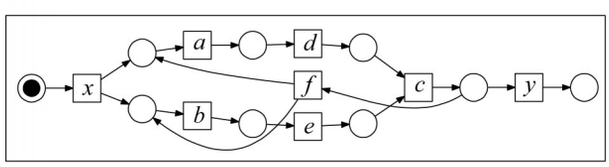

(a) 多路径选择　(b) 多任务并发　(c) 循环跳转　(d) 混合结构



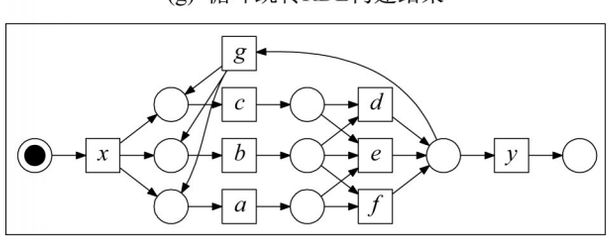

(e) 多路径选择RDL构建结果




(f) 多任务并发RDL构建结果



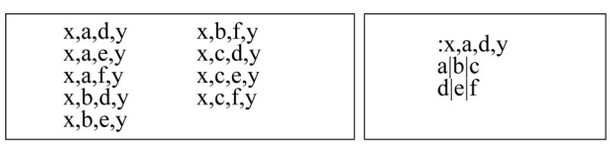

(g) 循环跳转RDL构建结果



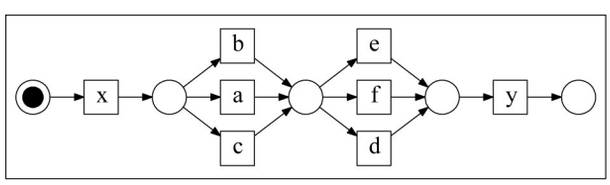

(h) 混合结构RDL构建结果

图5　典型工作流构建案例
Fig. 5　Typical workflow construction case

Sketch Miner方法将用户输入的事例轨迹片段组合成完整的事例轨迹，然后将其作为输入通过第三方的过程挖掘软件计算得到工作流模型，但是如果在流程中出现选择项或并发项较多，分支较复杂时，该方法的表达并不友好，且需要输入的文本量也将剧增。由于Sketch Miner轨迹输入依赖于相邻事件的跟随关系，当相邻事件可选择的数量均较多时，Sketch Miner需要多条语句来完成工作流描述，但采用RDLModel则简单得多，如图6所示的情况，两个环节各有3个可选路径，则此时Sketch Miner需要通过9条语句完成各种可能情况的描述，而RDLModel只用3条语句就可完成描述(为显示方便，在图6，7中对Sketch Miner DSL文本格式做了微调)。另外针对如图7所示的依赖条件较复杂的情况，若采用Sketch Miner描述该工作流，将不可避免需要进行分段记录，从而造成记录难度增加，记录产生的描述文本也因难以维护。相同情况下，RDLModel采用从整体到细节的逐层记录策略。首先对相邻并发关系活动进行记录，而后采用断开语句记录特定活动被触发的限制条件，从而实现对目标过程更优的描述，描述语句明显减少。

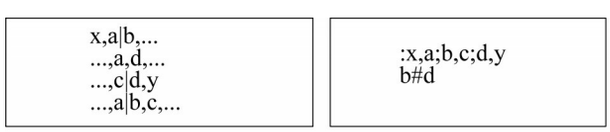



(a) Sketch Miner DSL　(b) RDL

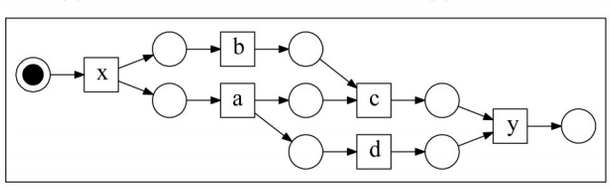



(c) 多分支工作流构建结果

图6　工作流描述文本对比
Fig. 6　Workflow description text comparison




(a) Sketch Miner DSL　(b) RDL




(c) 依赖条件较复杂的工作流构建结果

图7　工作流描述文本对比
Fig. 7　Workflow description text comparison

### 4.2 科学工作流应用案例

将本文提出的工作流构建方法用于基因测序高通量计算工作流的构建，该工作流采用DIVIS轻量级工作流引擎[32]。测序计算涉及较多的计算环节，对于不同的测序计算任务，要构建针对性的工作流以实现更优的计算效果。通过分层级和分环节的关系描述语句协同建模，系统基于RDL生成事例集并通过过程发现生成整体工作流。本案例将整体工作流分成包含5个环节的总体流程以及5个环节流程的2个层级组成。5个环节分别为预处理环节、比对环节、突变检测环节、标注环节、统计环节。总体流程由一个关系描述文本描述，各环节所包含的子环节由各自的关系描述文本描述。所涉及的流程建模结果如图8所示。

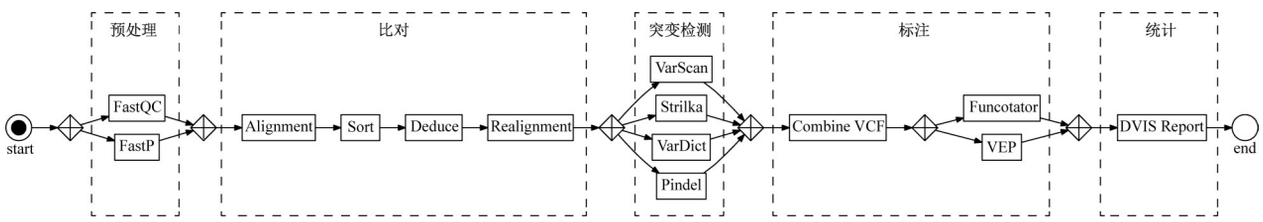



图8 由RDLModel构建生成的DIVIS基因测序工作流
Fig. 8 DIVIS gene sequencing workflow generated by RDLModel

由于采用了上述的工作流构建策略，工作流各层级及各环节的关系描述文本可分别由对该领域具有特别专长的技术人员维护。用户反馈该工具提高了工作流建模效率，可及时响应由于系统方法或支撑工具的变化所带来的工作流升级需求，从而提升了整体计算服务效率及效果。

## 5 结论

科学工作流的应用和管理是科学计算领域的重要问题。虽然工作流相关的工具软件层出不穷，但针对工作流构建的研究并未引起足够重视，相关研究工作并不深入，这也限制了科学工作流在规模和复杂度方面的发展。为了建设更具扩展性的工作流构建及维护系统，规范化工作流的生成和管理，让不同领域的专业人员协同完成高质量的科学工作流，提升科学计算实施人员的工作流管理效率，本文将过程挖掘方法引入到科学工作流的构建过程中，设计了一套全新的工作流描述语言以及相关的过程发现算法。通过启发式的流程片段收集以及关系描述语言RDL的灵活记录，工作流中的不同层级和不同环节的过程被方便地描述，通过文本转译生成流程关系数据，并采用过程发现算法生成健全的工作流网络模型。本文还介绍了交互式工作流构建工具RDLModel，一种基于RDL语法的交互式工作流构建及发现工具，在应用于多种形式的工作流构建及科学工作流实际应用案例中时，表现出较现有方法更优的构建效率。这种方法的新颖之处在于，它可以根据领域专业人员对工作流整体或局部的故事性描述，通过过程挖掘技术自动推导出控制流，并且当用户在界面中键入描述文本时，可以快速呈现出过程模型。而在传统方法中，工作流的发现只能在业务分析师的脑海中完成，然后再通过拖拽式图形工具来表现。

通过测试，也发现了现有方法的一些问题，比如由于α算法的计算量会随着活动数量的提升而显著增加，这将影响交互过程中模型构建结果的快速反馈，相应的算法优化工作有待开展。而且现有的交互工具功能较单一，配套辅助功能有必要进行更多的完善。

### 参考文献：

[1] Duncan Hull, Katy Wolstencroft, Robert Stevens, et al.




DOI: 10.16182/j.issn1004731x.joss.21-0151

Taverna: A Tool for Building and Running Workflows of Services[J]. Nucleic Acids Research (S0305-1048), 2006, 34(2): 729-732.

[2] Bertram Ludäscher, Ilkay Altintas, Chad Berkley, et al. Scientific Workflow Management and the Kepler System [J]. Concurrency and Computation: Practice and Experience (S1532-0626), 2006, 18(10): 1039-1065.

[3] Jalili V, Afgan E, Gu Q, et al. The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2020 Update[J]. Nucleic Acids Research (S0305-1048), 2020, 48(W1): 395-402.

[4] Hofmann, M, Klinkenberg R. RapidMiner: Data Mining Use Cases and Business Analytics Applications[M]. United Kingdom: Chapman and Hau/CRC, 2016: 343-372.

[5] Berthold, M, R. Cebron N, Dill F, et al. KNIME-the Konstanz Information Miner: Version 2.0 and Beyond[J]. AcM SIGKDD Explorations Newsletter (S1931-0145), 2009, 11(1): 26-31.

[6] Ivanchikj A, & Pautasso C. Sketching Process Models by Mining Participant Stories[C]// International Conference on Business Process Management. Vienna,Austria: BPM, 2019: 3-19.

[7] J Cook, A Wolf. Discovering Models of Software Processes From Event-Based Data[J]. ACM Transactions on Software Engineering and Methodology (S1557-7392), 1998, 7(3): 215-249.

[8] Murata, Tadao. Petri Nets: Properties, Analysis and Applications[J]. IEEE (S0018-9219), 1989, 77(4): 541-580.

[9] Van der Aalst, Wil MP. The Application of Petri Nets to Workflow Management[J]. Journal of Circuits, Systems, and Computers (S0218-1266), 1998, 8(1): 21-66.

[10] Peter Fettke. Business Process Modeling Notation[J]. Wirtschaftsinformatik und Angewandte Informatik (S0937-6429), 2008, 50(6): 504-507.

[11] Van der Aalst W. Causal Nets: A Modeling Language Tailored towards Process Discovery[C]// International Conference on Concurrency Theory. Aachen, Germany: Springer-Verlag, 2011: 28-42.

[12] La Rosa M, Reijers H A, Van Der Aalst W M, et al. APROMORE: an Advanced Process Model Repository [J]. Expert Systems With Applications (S0957-4174), 2011, 38(6): 7029-7040.

[13] Nadja Damij. Business Process Modelling Using Diagrammatic and Tabular Techniques[J]. Business Process Management Journal (S1463-7154), 2007, 13(1): 70-90.

[14] Ellson J, Gansner E, Koutsofios L, et al. Graphviz—Open Source Graph Drawing Tools[C]// International Symposium on Graph Drawing. Vienna, Austria: GD, 2001: 483-484.

[15] Robin Bergenthum, Jörg Desel, Sebastian Mauser, and Robert Lorenz. Construction of Process Models from Example Runs[J]. Transactions on Petri Nets and Other Models of Concurrency (S1867-7193), 2009, 2(1): 243-259.

[16] Dirk Fahland. Oclets-Scenario-Based Modeling with Petri Nets[C]// Int'l Conference on Applications and Theory of Petri Nets. Paris, France: LNCS, 2009: 223-241.

[17] Weijters A J, Van der Aalst W M. Rediscovering Workflow Models from Event-based Data using Little Thumb[J]. Integrated Computer-Aided Engineering (S1069-2509), 2003, 10(2): 151-162.

[18] Dixit P M, Verbeek H M W, Buijs J C, et al. Interactive Data-driven Process Model Construction[C]// International Conference on Conceptual Modeling, Xian, China: CCF, 2018: 251-265.

[19] A. Weijters, W. van der Aalst. Rediscovering workflow models from event-based data[C]// 11th Dutch-Belgian Conference of Machine Learning. Antwerp, Belgium: Benelearn, 2001: 93-100.

[20] J Herbst, D Karagiannis. An Inductive Approach to the Acquisition and Adaptation of Workflow Models[C]// IJCAI'99 Workshop on Intelligent Workflow and Process Management: The New Frontier for AI in Business. Stockhdm, Sweden, 1999: 52-57.

[21] R. Agrawal, D. Gunopulos, F. Leymann. Mining Process Models from Workflow logs[C]// 6th Int. Conf. Extending Database Technology (EDBT). Valencia, Spain: Springer, 1998: 467-483.

[22] W van der Aalst, T Weijters, L Maruster. Workflow Mining: Discovering Process Models from Event Logs[J]. IEEE Trans. Knowl. Data Eng. (S1041-4347), 2004, 16(1): 1128-1142.

[23] Q. Guo, L. Wen, J. Wang, et al. Mining Invisible Tasks in Non-Free-Choice Constructs[C]//International Conference on Business Process Management. Innsbruck, Australia 2015: 109-125.

[24] A Augusto, R Conforti, M Dumas, et al. Automated Discovery of Structured Process Models From Event Logs: The Discover-and-Structure Approach[J]. Data and Knowledge Engineering (S0169-023X), 2018, 117(1): 373-392.

[25] A Weijters, J Ribeiro. Flexible Heuristics Miner (FHM)[C]// IEEE Symposium on Computational Intelligence and Data Mining, Paris, France BETA, 2011: 310-317.

[26] F. Mannhardt, M. de Leoni, H. A. Reijers, et al. Data-Driven Process Discovery-Revealing Conditional Infrequent Behavior from Event Logs[C]// International Conference on Advanced Information Systems Engineering. Essen, Germany Springer, Cham, 2017: 545-560.

[27] S van Zelst, B van Dongen, W van der Aalst, et al. Discovering Workflow Nets Using Integer Linear Programming[J]. Computing (S0010-485X), 2017, 100 (5):1-28.

[28] Leemans S. J. J., Fahland D., Aalst W. M. P. van der. Discovering Block-Structured Process Models from Event Logs-a Constructive Approach[C]// PETRI NETS'13 34th International Conference on Application and Theory of Petri Nets and Concurrency, Milan, Italy: Springer-Verlag, 2013: 311-329.

[29] S. J. J. Leemans, D. Fahland, W. M. P. van der Aalst. Discovering Block-Structured Process Models from Event Logs Containing Infrequent Behaviour[C]// Business Process Management Workshops-BPM 2013 International Workshops. Beijing, China, 2014: 66-78.

[30] A K A de Medeiros, A J Weijters, W M van der Aalst. Genetic Process Mining: An Experimental Evaluation[J]. Data Mining and Knowledge Discovery (S1384-5810), 2007, 14(2): 245-304.

[31] J C Buijs B, F. van Dongen, W M van der Aalst.Quality Dimensions in Process Discovery: The Importance of Fitness, Precision, Generalization and Simplicity[J]. International Journal of Cooperative Information Systems (S0218-8430), 2014, 23(1): 100-144.

[32] 张裕, 牛北方.一种轻量级分布式作业管理系统的设计与实现[J]. 科研信息化技术与应用, 2019, 10(1): 31-37.
Zhang Yu, Niu Beifang. Design and Implementation of a Lightweight Distributed Job Management System[J]. E-science Technology & Application, 2019, 10(1): 31-37.




DOI: 10.16182/j.issn1004731x.joss.21-0151