日志诱导下的形态学片段流程聚类方法

孙书亚 方 欢 方贤文
(安徽理工大学数学与大数据学院 安徽 淮南 232001)

摘 要 在业务流程管理系统中,执行同一目的的任务流可能存在若干事件集的多种不同排列方式,对应在日志上则表现为很多日志存在着诸多变化,同时具有很多业务的共性特征。因此,如何提取日志行为的共性,将多个相似日志的流程进行聚类,实现提取流程簇业务系统的共性,对相似流程的业务融合具有积极意义。文中提出了一种基于日志的流程聚类方法,首先对日志中的低频事件进行过滤,利用日志形态学片段提取公共的高频片段,进而通过形式自动机将提取的公共高频片段转换为相似日志的聚类中心;然后,提出基于形态学片段的业务组合方法产生流程模型共性的频繁执行路径,将相似的等价类形态学片段进行业务组合,得到组合后的Petri网模型,即为流程簇的聚类中心;最后,通过一个实际的案例验证了所提方法的可行性和有效性。

关键词 形态学片段,流程聚类,流程组合,Petri网
中图法分类号 TP391.9 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.08.011

Log-induced Morphological Fragments Process Clustering Method

SUN Shu-ya FANG Huan FANG Xian-wen
(School of Mathematics and Big Data, Anhui University of Science & Technology, Huainan, Anhui 232001, China)

Abstract In the business process management system, there may be many different arrangements of several event sets in the task flow for performing the same purpose. Corresponding to the logs, it shows that many logs have many changes, but also have some common characteristics of many services. Therefore, extracting the commonality of the logs behavior and clustering multiple similar logs of the similar type of business system have positive significance for the business integration of similar processes. This paper proposed an approach of process clustering method. Firstly, low-frequency events are filtered out, and common high-frequency fragments from the morphological fragments in the log are extracted by automata. And then the extracted public high-frequency fragments are converted into clusters of similar logs through automation formal method. Then, a morphological fragment-based approach is proposed. A business combination algorithm is generated for those frequent execution paths of the commonality of the process model. By combining similar equivalent morphological fragments for business combination, a fused Petri net model is obtained. Finally, a practical case is proposed to verify the feasibility and validity of the proposed method.

Keywords Morphological fragments, Process clustering, Process combination, Petri net

业务流程管理已经越来越广泛地被应用于企业和组织的日常管理中,随着业务流程研究技术的发展,出现了诸多新的挑战,例如将数据科学中的一些数据思维融合到业务流程管理中,实现新的"数据+行为"业务流程管理框架[1]。类似于数据挖掘中的数据"聚类"(Data Clustering)思想,近些年涌现了许多从事件日志中挖掘系统模型的流程挖掘方法[2-5]。在业务流程簇(Process Families)中存在很多结构类似但是功能不同的业务系统模型集合,这些群体模型的分析和挖掘是目前研究的热点问题。目前研究尚未涉及从流程簇角度出发确定聚类中心的思想,因此本文的研究重点是将流程簇中的聚类中心作为流程簇中诸多相似的业务流程的代表,即将这些相似的业务流程(或流程片段)融合为一个新的业务流程(或流程片段)。

业务流程的建模与挖掘早已开始研究,特别是基于日志信息的过程挖掘方法取得了很多较完善的结论,其中Petri网是一种重要的业务流程形式化及建模工具,将Petri网用于流程片段的提取和挖掘是一种较为有效的方法[6-8],流程簇模型及其行为分析与研究大都与流程变体(Process Variant)[9-11]有关。文献[2]提出了一种通过分析事件日志来提取工作流模型的协同聚类方法,该方法打破了主要从事件日志中学习

到稿日期:2018-08-15 返修日期:2018-11-22 本文受国家自然科学基金项目(61472003,61272153,61340003,61402011,61572035),安徽省自然科学基金项目(1608085QF149),安徽省高校优秀青年人才基金项目(gxyqZD2018038),安徽省博士后基金项目(2018B288)资助。
孙书亚(1994-),女,硕士,主要研究方向为Petri网理论与应用,E-mail:2351982104@qq.com;方 欢(1982-),女,博士,副教授,主要研究方向为Petri网理论与应用、业务过程管理方法与应用,E-mail:fanghuan0307@163.com(通信作者);方贤文(1975-),男,博士,教授,主要研究方向为Petri网和可信软件。

单个工作流模型的方法的局限性。同时将用户学习模式UBP建模为序列上的概率分布,允许计算UBP和序列之间的相似性,共同聚集用户和序列,以生成两种类型的聚类:用户群集和序列聚类,用于对相同工作流模型实例的序列进行分组。文献[4]介绍了一种对流程变体簇建模的方法,用于解决对变体进行单独建模会引起共享公共部分的模型数量增加,从而导致整体模型冗余和不一致的问题。文献[7]基于Bung本体论评估了现有的各种流程片段定义,提出了一种支持聚类的算法以支持流程片段识别的可重用性和灵活性。文献[12]旨在发现模型不同变体之间的共同点和差异,根据事件日志集合处理模型,通过实例和真实事件评估所提方法的优良性。文献[13]引入了基于面向对象方法动态修改正在运行的工作流实例的体系结构,并且引入了案例处理作为基于数据依赖性而不是流程结构的流程制定技术,利用事件日志作为输入配置流程片段解决了在流程设计中引入模型的复杂性。文献[14]提出了一种从日志中挖掘公共流程片段的算法。文献[15]提出了一种在特定活动周围合并流程片段,构建合并片段的算法。该合并片段作为可配置的子流程呈现,并且可由业务流程设计者配置以创建业务流程变体。文献[16]提出了一种被称为Log to Model Explorer的集成工具,用于对日志进行过滤、聚类及标签细化,帮助用户以交互方式迭代地探索和预处理日志,使他们能够发现更有意义的流程模型。文献[17]讨论了从共享公共片段的流程模型集合中构建整合业务流程模型的问题,提出了一种用于计算合并模型的算法。文献[18]提出了一种合并业务流程片段的方法,以便于在业务流程设计中重用片段,该方法依赖于所谓的邻接矩阵,所提出的方法已经在一系列片段上实施和测试,并得到了有效的结果。

从已有的研究工作和成果不难看出,流程片段在大型的跨组织流程的设计中发挥着重要的作用,如何提取和配置流程片段将是未来模型设计的关键。目前针对流程簇中的流程片段的相关研究大多集中在两个方面:1)对日志中片段的差异进行识别,从而提取可变的流程片段,通过日志集合配置流程模型;2)以可配置的参考模型为输入,构建可配置的流程片段进行业务流程设计。

然而,在流程模型的配置和设计过程中,以上两个方面的研究均未涉及日志中片段的相似性或共性特征[19-20],因此提炼多个相似流程或者日志的共性,计算多个相似流程或日志的聚类中心,这种"数据+行为"的业务流程管理方法具有重要的应用价值,例如在大型企业或者组织中需要解决管理者如何选择的问题:决策者要对大量的数据信息进行采集、分析和决策,而面对庞大的数据存储则需要进行分析整理,保留多个决策方案的共性,提取不同方案执行过程中频繁出现的相似行为,并将这些频繁出现的相似行为进行融合和聚类,凝练出相似行为的聚类中心。这种行为的聚类中心思想类似于数据的聚类中心,然而行为的聚类中心主要从行为关系的角度出发。

基于此,本文通过分析同一组相似日志中的片段之间的关系,以日志片段的相似性研究为着手点,以Petri网为形式化建模工具,获取行为相似的等价形态学片段。进而,提出从事件日志集合中提取公共形态学片段的方法,计算日志中出现的流程片段并保留高频出现的公共片段,以产生相似日志的聚类中心,通过自动机理论对片段进行组合,获取具有唯一执行路径的Petri网模型。

本文第1节主要定义了基本概念;第2节对形态学片段的提取及检测算法进行了详细的阐述;第3节在第2节的基础上提出了片段组合的算法以实现日志片段组合;第4节利用一个实际案例验证所提方法的实用性和有效性;第5节将本文工作与已有研究工作进行分析和比较;最后总结全文,并对未来工作进行展望。

1 基本定义

限于篇幅,有关Petri网的概念及结构的定义在此不做赘述,具体内容可以参考文献[19]。

定义1[7](事件日志) 若T是任务集,则称$\sigma\in T^*$是一条事件轨迹,$W\subseteq T^*$为一个事件日志。

定义2[7](流程片段) 一个流程片段F可表示为$F(A,G,R,s,e)$,F被定义为定向图当且仅当满足条件:

(1)$R\subseteq(A\times G)\cup(G\times A)\cup(G\times G)$;

(2)$|A|\geqslant 1$;

(3)$\forall t\in(A\cup G)\ \exists v_0=s,v_1,v_2,\cdots,v_k=t((v_{i-1},v_k)\in R,1\leqslant i\leqslant k)$;

(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)$\exists t\in(A\cup G)((v_t,v_i)\in R\vee(v_k,v_t)\in R)\vee((v_j,v_{j+1})\in R,i\leqslant j\leqslant k)$。

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3[14](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,则称之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3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。下面举例说明流程片段及等价形态学片段两种片段的区别和联系。

图1给出了4种结构不同但操作相同的片段,图1(a)—图1(d)分别是4种结构不同的流程片段,此类片段的特征为:开始和结束节点相同,并且具有相同的中间节点集。因此,这4个片段称为等价的形态学片段。

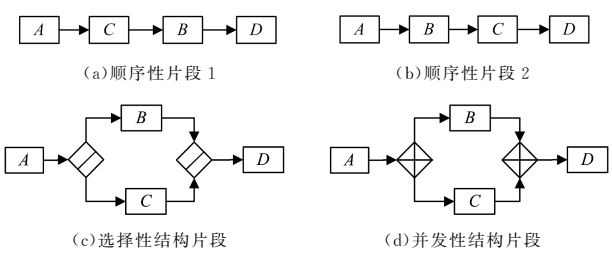


图1 结构不同的4种等价形态学片段

Fig. 1 Four equivalent morphological fragments with different structures

2　等价形态学片段的提取与检测

2.1　等价形态学片段的提取及检测算法

本文基于事件日志提取形态学片段，为了减少日志噪音，首先过滤事件日志中的非频繁事件，然后计算产生频率较高的流程片段，最后对日志中的等价形态学片段进行检测和提取，具体如算法 1 所示。

算法 1　等价形态学片段提取及检测算法（Equivalence Morphological Fragment Extraction and Detection Algorithm，EMFEDA）

输入：事件日志集合 TS^i

输出：等价的形态学片段集合 A

Step1　对于数据集 i，计算不同的迹 t 的发生频率 LF_t^i；

Step2　在每个数据集 i 中，比较每条迹的发生频率 LF_t^i，对发生频率 $LF_t^i<\beta(\beta=0.05)$ 的迹进行过滤，反之予以保留；

Step3　对于 TS^i 中的每一个迹 t，产生所有长度为 L 的片段 LG_t^i，$2\leqslant L\leqslant t_{max}$；

Step4　对于每一个属于 LG_L^i 中的片段 L，利用下式对其进行数值表示：

$$L=\mu\times A_1+\left(\sum_{j\neq 1,j\neq 0}A_j\right)+\omega\times A_0$$

其中，μ,ω 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。

Step5　对于每一个 L，在所有具有相同数值的 LG_L^i 中计算它的全局频率 GF_L；

Step6　返回所有 $GF_L>1$ 的等价形态学片段。

下面对上述算法的步骤给出具体分析。

2.1.1　过滤非频繁事件迹

流程模型可能包含成千上万个事件迹，我们把流程模型 i 包含的迹的有限非空集合表示为 TS^i。首先，对于出现在事件日志中的不同的迹，计算其发生的频率，迹 t 的全局频率表示为 LF_t^i，设定阈值 $\beta=0.05$，对发生频率小于阈值 $\beta(LF_t^i<\beta)$ 的迹进行过滤，保留在事件日志中出现频率较高的迹。接下来计算每条迹发生的全局频率，事件日志总的实例数为 2648，可以很容易地计算每条迹的 LF_t^i，$LF_t^i(EHABCDFGABCD)=\frac{650}{2648}\approx 0.2455$，$LF_t^i(EHABCDFG)=\frac{328}{2648}\approx 0.1239$，……依次进行计算可得迹 $HEABCDFG$ 的全局频率 $LF_t^i\approx 0.045<\beta$ 并对其进行过滤。

表 1 列出了原始事件日志集，表 2 列出了过滤后的事件日志。

表 1　一组事件日志

Table 1　A set of event log

实例数	事件轨迹
650	EHABCDFGABCD
328	EHABCDG
700	HEACBDGFABCD
230	EHBCDAGFBADC
120	HEABCDFG
215	HEBDACFG
405	HEBDCAGFABCD

表 2　过滤后的事件日志

Table 2　Filtered event log

实例数	事件轨迹
650	EHABCDFGABCD
328	EHABCDG
700	HEACBDGFABCD
230	EHBCDAGFBADC
215	HEBDACFG
405	HEBDCAGFABCD

2.1.2　提取频繁的形态学片段

TS^i 中的每一个迹都可以产生尺寸大小为 $L(2\leqslant L\leqslant\|t_{max}\|)$ 的片段，L 的最大取值为最长的迹 t_{max} 的长度，最小取值为 2。以日志 $HEABCDFG$ 为例，令 $L=4$ 产生所有的片段为：$HEAB$，$EABC$，$ABCD$，$BCDF$，$CDFG$。对表 2 中的个别事件日志产生长度大小为 $L=4$ 的片段，如下所示：

$$EHABCDFGABCD\begin{cases}EHAB\\HABC\\ABCD\\BCDF\\CDFG\\\underline{DFGA}\\FGAB\\GABC\\ABCD\end{cases},\ HEACBDGFABCD\begin{cases}HEAC\\EACB\\ACBD\\CBDG\\BDGF\\\underline{DGFA}\\GFAB\\FABC\\ABCD\end{cases}$$

由给出的部分事件日志的片段可得：片段 $ABCD$ 和片段 $ACBD$ 为不同迹中相同位置的等价形态学片段，片段 $DFGA$ 和片段 $DGFA$ 同样为等价的形态学片段，本文用不同的形式对不同类别的形态学片段进行标记。在对日志的相似性进行聚类的过程中，提取同等位置执行相同任务时出现频率较高的片段，而对于不同的事件序列，若同时出现对应位置的等价形态学片段，则可选择保留其一用于业务流程组合。

2.1.3　形态学片段检测

为了进一步识别并检测形态学片段，本文提出了一种快速而精确的方法。首先，假设在流程执行中有 8 个活动，分别是 A,B,C,D,E,F,G 以及 H，把每一个活动映射为不同的数值，用 10 的幂次进行表示，如图 2 所示。

A	B	C	D	E	F	G	H
↓	↓	↓	↓	↓	↓	↓	↓
10^7	10^6	10^5	10^4	10^3	10^2	10^1	10^0

图 2　活动的数值表示

Fig. 2　Numerical representation of activities

利用式(1)可以对每一条迹 t 分配独一无二的数值表示：

$$Value(t)=\mu\times A_I+\left(\sum_{j\neq I,j\neq O}A_j\right)+\omega\times A_O \tag{1}$$

其中，A_I 是迹中第一个活动的数值表示；A_O 是迹中最后一个活动的数值表示；$A_j(j\neq I,j\neq O)$ 是迹中除了第一个和最后一个活动外，其他活动的数值表示，系数固定为 1。利用式(1)，每个活动有唯一的映射数值，因此能够对每一个片段产生唯一的数值表示。对于第一个和最后一个活动可设置取值在 2～9 之间的不同系数，假设用字母 μ 和 ω 分别表示第一个和最后一个活动的系数，令 $\mu=2,\omega=3$（μ 和 ω 是不同于其他活动系数的互异数值）。具有相同数值的片段为等价的形

态学片段，片段 $ABCD$ 和片段 $ACBD$ 的数值表示相同，称为等价形态学片段，而片段 $DFGA$ 与片段 $DAGF$ 相比，数值表示不等，因为其对应的最后一个活动不相同。具体的数值表示如表 3 所列。

表 3 迹的数值表示

Table 3 Numerical representation of trace

迹	映射计算	数值表示
$DFGA$	$2\times10^4+10^2+10^1+3\times10^7$	30020110
$ABCD$	$2\times10^7+10^6+10^5+3\times10^4$	21130000
$ACBD$	$2\times10^7+10^5+10^6+3\times10^4$	21130000
$DAGF$	$2\times10^4+10^7+10^1+3\times10^2$	10020310

3 流程组合后的聚类中心计算

利用从日志中提取的高频片段进行组合，得到包含等价形态学片段的聚类中心。将这个聚类的行为中心进行业务组合以获取一种可能的执行流程，即组合后的 Petri 网模型，从而降低相似流程的共性设计复杂度。具体算法如算法 2 所示。

算法 2 片段组合算法(Fragment Combination Algorithm, FCA)

输入：事件日志集合 TS

输出：组合 Petri 网模型 M(聚类中心)

Step1 对于 TS 中的每一个迹 T，产生所有长度为 L($2\leqslant L\leqslant T_{max}$)的片段。

Step2 取 L=4，计算得出每条迹 T_i($i=1,2,\cdots,n$)对应的片段集 FS_i($i=1,2,\cdots,n$)，对于不同迹中同等位置的片段保留出现频率最高的片段 F_{max}。若片段集 FS_i($i=1,2,\cdots,n$)中 F_i 与片段 F_j 为等价的形态学片段，且频率和较高，则保留 F_i 或 F_j。

Step3 利用自动机理论 M=(Q,Σ,δ,q_0,F)，记录片段的状态转换过程；对保留的高频片段进行业务组合，从而得到具有唯一执行路径的 Petri 网模型 M。

根据片段组合算法，以一组有限的日志为例，从中提取频繁出现的片段设计一种可能的流程组合。

如表 4 所列，从每条迹的第一个活动开始逐渐产生大小为 $L=4$ 的流程片段，按迹的顺序对应的片段依次如表 5 所列。

表 4 有限的事件日志集

Table 4 Limited event log set

变迁序列	事件轨迹	变迁序列	事件轨迹
$t1$	$EAHBCDFGABCD$	$t5$	$HBEACDGFABCD$
$t2$	$HEABCDFG$	$t6$	$HEABCDFG$
$t3$	$HEABCDGFABCD$	$t7$	$EHABCDFGABCD$
$t4$	$EHABCDGFABCD$	$t8$	$HAEBCDGFABCD$

表 5 $L=4$ 的流程片段

Table 5 Process fragments of $L=4$

$EAHB$	$AHBC$	$HBCD$	$BCDF$	$CDFG$	$DFGA$	$FGAB$	$GABC$	$ABCD$
$HEAB$	$EABC$	$ABCD$	$BCDF$	$CDFG$				
$HEAB$	$EABC$	$ABCD$	$BCDG$	$CDGF$	$DGFA$	$GFAB$	$FABC$	$ABCD$
$EHAB$	$HABC$	$ABCD$	$BCDG$	$CDGF$	$DGFA$	$GFAB$	$FABC$	$ABCD$
$HBEA$	$BEAC$	$ABCD$	$BCDF$	$CDGF$	$DGFA$	$GFAB$	$FABC$	$ABCD$
$HEAB$	$EABC$	$ABCD$	$BCDF$	$CDFG$				
$EHAB$	$HABC$	$ABCD$	$BCDF$	$CDFG$	$DFGA$	$FGAB$	$GABC$	$ABCD$
$HAEB$	$AEBC$	$ABCD$	$BCDG$	$CDGF$	$DGFA$	$GFAB$	$FABC$	$ABCD$
$HEAB$	$EABC$	$ABCD$	$BCDF$	$CDFG$	$DFGA$	$GFAB$	$FABC$	$ABCD$
$HAEB$			$BCDG$	$CDGF$	$DGFA$			

表 5 中最后一行片段为对应列中出现频率最高的片段 F_{max}，其余每行代表迹 t_i($i=1,2,\cdots,8$)中 $L=4$ 的所有片段，每一列代表不同的事件序列对应位置的 $L=4$ 的流程片段。用字母 F_i($i=1,2,\cdots,9$)表示对应的片段，如下所示：

$$\begin{matrix} F_1 & F_2 & F_3 & F_4 & F_5 & F_6 & F_7 & F_8 & F_9 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ \begin{bmatrix} HEAB \\ HAEB \end{bmatrix} & [EABC] & [ABCD] & \begin{bmatrix} BCDF \\ BCDG \end{bmatrix} & \begin{bmatrix} CDFG \\ CDGF \end{bmatrix} & \begin{bmatrix} DFGA \\ DGFA \end{bmatrix} & [GFAB] & [FABC] & [ABCD] \end{matrix}$$

有穷状态自动机是一种重要的计算机形式语言，能够对信号序列进行逻辑处理，能与外界交换信息并改变动作，因此本文利用自动机理论对片段状态转换进行描述，用来表示活动的执行路径，并比较通过日志聚类所获取的频繁执行路径与原事件日志的变化距离。下面首先给出有穷状态自动机及基于日志的变化操作的相关概念。

定义 4[21](有穷状态自动机) M 是一个五元组：$M=(Q,\Sigma,\delta,q_0,F)$，其中，状态的非空有穷集合为 Q。$\forall q\in Q$，q 称为 M 的一个状态；输入字母表为 Σ。输入字符串都是 Σ 上的字符串。状态转移函数为 δ，有时称作状态转换函数或者移动函数，$\delta:Q\times\Sigma\to Q$。对于 $\forall(q,a)\in Q\times\Sigma$，$\delta(q,a)=p$ 表示 M 在状态 q 读入字符 a，将状态变成 p，并将读头向右移动一个带方格而指向输入字符的下一个字符。q_0 为 M 的开始状态，也称作初始状态或者启动状态，$q_0\in Q$。F 为 M 的终止状态集合，$F\subseteq Q$。$\forall q\in F$，q 称为 M 的终止状态。

定义 5(变化操作) W 为一个事件日志，事件迹 $\sigma\in W$，日志中的活动集合记为 S，下面针对日志中的活动集合及活动顺序的变化，定义 3 种变化操作。

$\forall a_i\in S,\sigma_i\in W(i=1,2,\cdots,n)$

$Move$：$Move(\sigma_i,a_1,a_2,a_3)$。移动操作是指在迹 σ_i 中把活动 a_1 从当前位置移动到活动 a_2 之后和 a_3 之前。

$Delete$：$Delete(\sigma_i,a_i)$。删除操作是指把活动 a_i 从迹 σ_i 中删除。

$Insert$：$Insert(\sigma_i,a_1,a_2,a_3)$。插入操作是指在迹 σ_i 中，把活动 a_1 插入到 a_2 和 a_3 之间。

把 3 种变化操作记为：$C=(Move,Delete,Insert)$。

如果 $\sigma_{i-1}[\Delta>\sigma_i$ 成立(Δ 指应用变化操作的种类)，那么 σ_{i-1} 与 σ_i 之间的距离 $d(\sigma_{i-1},\sigma_i)=|\Delta|$。

利用自动机理论对片段进行组合，首先给出以下形式化表示：

(1)$Q=\{F_1,F_2,F_3,F_{4_1},F_{4_2},F_{5_1},F_{5_2},F_6,F_7,F_8,F_9\}$；

(2)$\Sigma=\{A,B,C,D,E,F,G,H\}$；

(3)$\delta:Q\times\Sigma\to Q$；

(4)$q_0=F_1$；

(5) $F=F_9$。

片段转换模式如图 3 所示。图 3(a)中，系统在状态 F_1 时启动，用箭头 S 表示，当读入字符 C 时，转到状态片段 F_2，状态 F_2 读入字符 D，转到状态片段 F_3…，依次进行字符的读入和状态的转换，直到系统到达最终状态 F_9，用双圈顶点进行表示。根据状态转换过程，图 3(b)记录了活动执行的顺序路径，从图中可以看出总共有 4 条频繁路径的执行。由于片段 $HEAB$ 和 $HAEB$ 为等价的形态学片段，则总共包含两条频繁执行的路径，即：$HEABCDFGABCD$，$HAEBCDGFABCD$，转化为 Petri 网语言如图 4 所示。

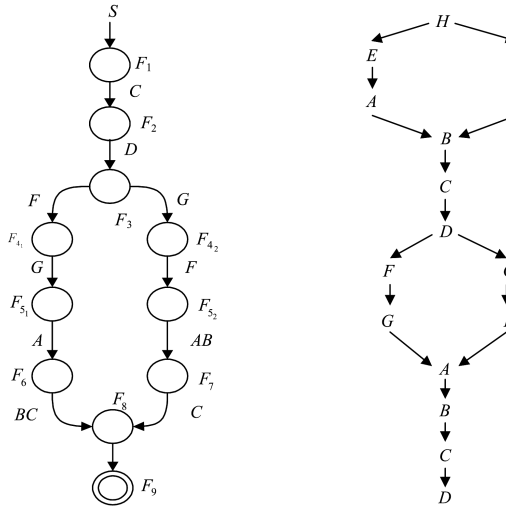
记路径 $T=HEABCDFGABCD$，下面通过比较 T 与迹 t_i ($i=1,2,\cdots,8$)(见表 4)之间的距离来度量本文所获取的执行迹与原事件日志序列的相似性，所对应的变化操作如下：

$T\rightarrow t_1$：$\Delta=[Move(T,H,A,B)]$，$d(T,t_1)=1$；

$T\rightarrow t_2$：$\Delta=[Delete(T,ABCD)]$，$d(T,t_2)=1$；

⋮

类似地，可以计算 $d(T,t_3)=1$；$d(T,t_4)=2$；$d(T,t_5)=2$；$d(T,t_6)=1$；$d(T,t_7)=1$；$d(T,t_8)=2$。可以明显地看出，组合后的聚类中心与原先日志片段模型的距离 $d(T,t_i)\leqslant 2$，这进一步表明通过片段组合所获取的执行迹能够很好地对原事件日志进行聚类，保留其相似性。路径 $HAEBCDGFABCD$ 同理。



(a)自动机下的片段转换模式　　(b)活动执行路径

图 3　片段转换模式及活动执行路径

Fig. 3　Fragment transformation pattern and activity execution paths

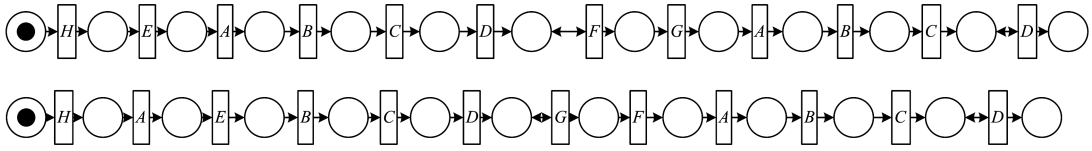


图 4　Petri 网语言转化图

Fig. 4　Petri net conversion chart

4　案例分析

图 5 给出了实际公司领导或者大型组织决策者在面对较复杂的任务时的决策流程。首先明确决策主题，然后采集信息并对获取的数据信息进行分析整理，最后确定最终的决策方案。在这一过程中，最重要的是决策者如何从众多决策参考中筛选出有用的信息来确定最佳方案。这就需要采取求同存异、共性保留的原则，提取不同方案中执行相同步骤的频繁出现的相似行为进行重新配置组合，从而得到新的决策方案。

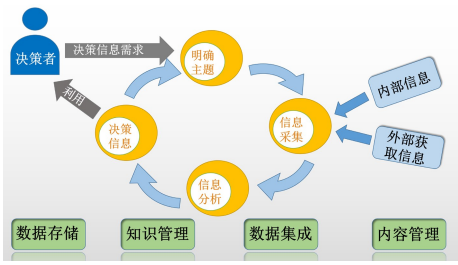


图 5　组织决策图

Fig. 5　Organizational decision picture

采用本文的流程组合聚类方法对决策图(见图 5)进行转化，转化过程如图 6 所示。

图 6 将图 5 中决策者在确定决策方案中的每一步的实施步骤转化为本文提出的流程聚类方法中的实施步骤，本文提出的日志聚类流程设计的方法有利于决策者对大量复杂的信息进行收集、整理，利用出现频数较高的形态学片段信息进行业务融合，获取新的组合模型。

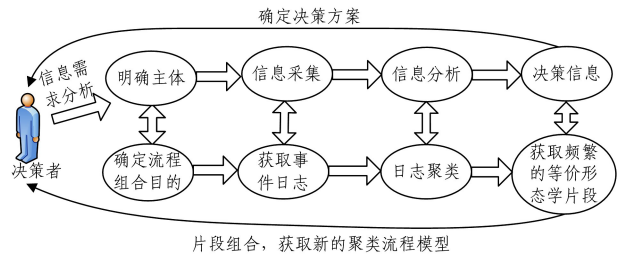


图 6　决策流程转化过程

Fig. 6　Decision process transformation process

下面以文献[19]中的公司采购为例，具体介绍本文所提方法在实际中的应用。首先给出在实际采购过程中的事件日志集，利用从日志中提取的等价形态学片段，通过自动机形式化语言对片段进行组合从而获取频繁执行的路径，即最终采购方案执行图，以使公司或组织在面对多决策选择时，能够快速并准确地做出决定。各字母所表示的内容如下：A 表示采购中心采购；C 表示选择供应商；D 表示公开招标；E 表示制作标书；F 表示发布公告；G 表示审核招标文件；H 表示发送招标文件；I 表示谈判准备；J 表示开标、唱标、评标；K 表示确认招标文件；L 表示询价采购；M 表示成立询价小组；N 表

示随机抽取三家有资质的供应商；O 表示询价；P 表示竞争性谈判；R 表示成立谈判小组；S 表示制作谈判文件；T 表示邀请三家以上供应商；U 表示谈判；V 表示审阅报价文件；W 表示技术谈判；X 表示商务谈判；Y 表示最终报价；Z 表示定价；a 表示发布中标公告；b 表示签发中标通知书；c 表示签订合同；d 表示邀请招标；e 表示邀请 3 家有资质的供应商；m 表示发出投标邀请。

采购公司在采购过程中存在多种采购方案，我们列举其中部分事件日志进行分析，利用事件日志中的每一条迹产生所有大小为 $L=4$ 的流程片段，根据等价形态学片段的等价性及自动机理论对片段进行组合产生唯一执行路径即确定广泛被采纳的采购方案。按照自上往下的顺序，依次计算迹对应位置产生的 $L=4$ 的流程片段，并对片段出现的频数进行标记，表格最后一行为相应位置出现频数最高的片段，若片段为等价的形态学片段且频数之和最高，将对其进行保留，在接下来的流程组合中选择等价片段之一作为可配置的流程片段组合流程模型，具体片段的产生如表 7 所列。其中，括号内的数字表示片段产生的频数。

表 6 某公司采购系统的事件日志

Table 6 Event log of a company purchasing system

变迁序列	事件轨迹	变迁序列	事件轨迹
$t1$	$ACLMNOVXYZabc$	$t7$	$ACIPTUVWZYabc$
$t2$	$ACLMNOVWYZbac$	$t8$	$APICSVUYWZbac$
$t3$	$ACPITUVWYZabc$	$t9$	$ACdemIJK$
$t4$	$ACPITUVXYZbac$	$t10$	$ACdemGHK$
$t5$	$ACPITUVWYZabc$	$t11$	$ACDEFGJK$
$t6$	$ACPIRUVWYZabc$	$t12$	$ACDEFGHK$

表 7 表 6 中事件日志对应的过程片段及相应频数

Table 7 Process fragments corresponding to event log and corresponding frequency

$ACLM(2)$	$CLMN(2)$	$LMNO(2)$	$MNOV(2)$	$NOVX(1)$	$OVXY(1)$	$VXYZ(2)$	$XYZa(2)$	$YZab(4)$	$Zabc(4)$
$ACPI(4)$	$CPIT(3)$	$PITU(3)$	$ITUV(3)$	$NOVW(1)$	$OVWY(1)$	$VWYZ(4)$	$WYZa(4)$	$YZba(2)$	$Zbac(3)$
$ACIP(1)$	$CPIR(1)$	$PIRU(1)$	$IRUV(1)$	$TUVW(3)$	$UVWY(3)$	$VWZY(1)$	$WZYa(1)$	$ZYab(1)$	$Yabc(1)$
$APIC(1)$	$CIPT(1)$	$IPTU(1)$	$PTUV(1)$	$TUVX(1)$	$UVXY(1)$	$UYWZ(1)$	$YWZb(1)$	$WZba(1)$	*
$ACde(2)$	$PICS(1)$	$ICSV(1)$	$CSVU(1)$	$RUVW(1)$	$UVWZ(1)$	*	*	*	*
$ACDE(2)$	$Cdem(2)$	$demI(1)$	$emIJ(1)$	$SVUY(1)$	$VUYW(1)$	*	*	*	*
*	$CDEF(2)$	$demG(1)$	$emGH(1)$	$mIJK(1)$	*	*	*	*	*
*	*	$DEFG(2)$	$EFGJ(1)$	$mGHK(1)$	*	*	*	*	*
*	*	*	$EFGH(1)$	$FGJK(1)$	*	*	*	*	*
*	*	*		$FGHK(1)$	*	*	*	*	*
$ACPI(4)$	$CPIT+CIPT(4)$	$PITU(3)$	$ITUV(3)$	$TUVW(3)$	$UVWY(3)$	$VWYZ(4)$	$WYZa+WZYa(5)$	$YZab(4)$	$Zabc+Zbac(7)$

根据算法 1 和算法 2，迹是按照自左往右的顺序依次产生 $L=4$ 的片段，对每条迹相同位置产生的流程片段进行比较。接着利用自动机形式化语言对片段组合状态进行表示。

图 7 给出了在采购流程中片段转换的过程，通过字符的读入实现状态的转换即采购步骤的执行，可以计算得出采购执行的路径：$T=ACPITUVWYZabc$。在路径 T 中每一个活动代表一个事件的发生，用变迁进行表示。将其转化为 Petri 网建模语言，如图 8 所示。

在所有的采购方案中，我们所获取的执行次数最多，广泛被采纳的方案如图 8 所示，即 A(采购中心采购)→C(选择供应商)→P(竞争性谈判)→I(谈判准备)→T(邀请三家以上供应商)→U(谈判)→V(审阅报价文件)→W(技术谈判)→Y(最终报价)→Z(定价)→a(发布中标公告)→b(签发中标通知书)→c(签订合同)。

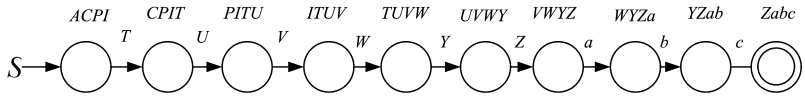


图 7 采购流程片段转换过程

Fig. 7 Purchasing process fragment transformation process

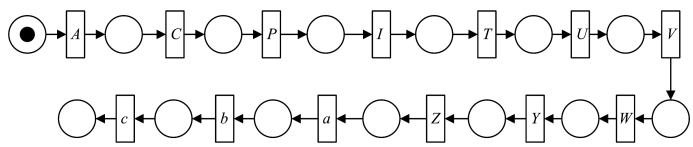


图 8 采购流程执行路径

Fig. 8 Purchasing process execution path

5 分析与比较

本节将本文的研究工作与已有研究进行分析比较，重点比较已有的流程组合的相关研究与本文提出的基于形态学片段的流程聚类方法，各种方法的汇总如表 8 所列。

从表 8 中可以看出，现有关于模型挖掘配置的方法主要从事件日志及流程变体两个方面来挖掘有用信息，以获取可配置的流程模型。而本文所提出的方法主要对日志进行聚类，获取可配置的等价形态学片段，组合出具有唯一执行路径的 Petri 网模型。

表 8　流程聚类方法总结

Table 8　Summary of process clustering methods

方法 \ 特点	前提条件	运用方法	结果分析	特点及局限
从事件日志中对可配置的流程模型进行流程设计[12]	事件日志集合	ETM 遗传算法的扩展	获得能够描述流程变体而非特定流程的模型	不能通过探索现有流程的使用来辅助流程的配置
从事件日志中挖掘可配置的流程片段，组合流程模型[17]	可配置的过程片段	动态指南驱动方法、基于频率的后缀树方法	对真实的业务流程进行设计	基于频率行为设计流程，忽略了其他有趣信息，如绩效信息等；未讨论配置片段的行为语义及语法正确性
从流程变体中挖掘可配置流程模型[18]	可配置的流程变体	提出将多个流程变体合并到一个流程模型的合并算法	(1)对所有原始模型的行为进行归类；(2)确保追溯到每个元素的起源；(3)通过配置和个性化来导出任意的输入模型	输入流程变体过大会导致配置模型的管理变得复杂
从日志中挖掘形态学片段组合模型(本文方法)	等价的形态学片段	日志聚类，流程组合	将相似的等价形态学片段进行业务融合，产生流程模型共性的频繁执行路径	有效地保留了日志共性，为业务融合奠定了基础

结束语　大型的相似复杂的业务流程不利于组织者对事件进行决策和筛选，本文的工作是对操作相同的高频等价形态学片段进行提取，找出事件日志中的片段相似性，提取公共的片段进行组合，从而得到相似流程模型的聚类中心，即相似流程模型中的共性特征，进而通过过滤非频繁行为得到频繁执行迹，即高频出现的执行路径。所获取的执行方案能够满足不同利益相关者的需求，并且最大可能地覆盖相似流程的共性特征。本文提出的算法可以嵌入到 ProM 软件来实现大型日志案例的处理，限于篇幅这部分的工作将另文阐述。

未来的工作重点是通过给定约束集利用流程片段对模型进行组合优化，尝试提取和使用事件日志中诸如成本及复杂性的特征，根据管理者要求或者个人喜好自动构造最佳的可配置的流程模型。

参 考 文 献

[1] AALST W M P V D. Process Mining - Data Science in Action (2nd edn)[M]. Springer, Heidelberg, 2016.

[2] LIU X, DING C. Learning Workflow Models from Event Logs Using Co-clustering[J]. International Journal of Web Services Research, 2013, 10(3): 42-59.

[3] LEONI M D, AALST W M P V D, DEES M. A general process mining framework for correlating, predicting and clustering dynamic behavior based on event logs[J]. Information Systems, 2016, 56(C): 235-257.

[4] MILANI F, DUMAS M, AHMED N, et al. Modelling families of business process variants: A decomposition driven method[J]. Information Systems, 2016, 56: 55-72.

[5] LI C, REICHERT M, WOMBACHER A. Discovering process reference models from process variants using clustering techniques[J]. Centre for Telematics & Information Technology University of Twente, 2018, 16(5): 1-30.

[6] WESKE M. Business Process Management: Concepts, Languages, Architectures[M]. Springer-Verlag New York, Inc. 2007.

[7] POURMASOUMI A, KAHANI M, BAGHERI E. Mining variable fragments from process event logs[J]. Information Systems Frontiers, 2017, 19(6): 1-21.

[8] MA H, TANG Y, WU L K. Model update method in process incremental mining[J]. Computer Science, 2009, 36(5): 154-157.

[9] BOLT A, LEONI M D, AALST W M P V D. Process Variant Comparison: Using Event Logs to Detect Differences in Behavior and Business Rules [J]. Information Systems Frontiers, 2018, 74(1): 53-66.

[10] DÖHRING M, REIJERS H A, SMIRNOV S. Configuration vs. adaptation for business process variant maintenance: An empirical study[J]. Information Systems, 2014, 39(1): 108-133.

[11] BUIJS J C A M, REIJERS H A. Comparing Business Process Variants Using Models and Event Logs[M]// Enterprise, Business-Process and Information Systems Modeling. Springer Berlin Heidelberg, 2014: 154-168.

[12] BUIJS J, DONGEN B, AALST W. Mining Configurable Process Models from Collections of Event Logs[C]// International Conference on Business Process Management. Springer-Verlag, 2013: 33-48.

[13] ASSY N, GAALOUL W, DEFUDE B. Mining Configurable Process Fragments for Business Process Design[M]// Advancing the Impact of Design Science: Moving from Theory to Practice. Springer International Publishing, 2014: 209-224.

[14] HASANKIYADEH A, KAHANI M, BAGHERI E, et al. Mining common morphological fragments from process event logs [C]// International Conference on Computer Science and Software Engineering. IBM Corp, 2014: 179-191.

[15] ASSY N, CHAN N, GAALOUL W, et al. Deriving configurable fragments for process design[J]. International Journal of Business Process Integration & Management, 2014, 7(1): 2-21.

[16] LU X X, FAHLAND D D, WIL V D A W. Interactively exploring logs and mining models with clustering, filtering, and relabeling[C]// Proceedings of the BPM 2016 Tool Demonstration TRACK. 2016.

[17] ASSY N, CHAN N, GAALOUL W. Assisting Business Process Design with Configurable Process Fragments[C]// IEEE International Conference on Services Computing. IEEE Computer Society, 2013: 535-542.

[18] DERGUECH W, BHIRI S. Merging Business Process Variants [C]// Business Information Systems, International Conference (Bis 2011). Poznan, Poland, DBLP, 2011: 86-97.

[19] 方贤文. Petri 网行为轮廓理论及其应用[M]. 上海：上海交通大学出版社，2017：39-40.

[20] ZEMNI M A, HADJ-ALOUANE N B, MAMMAR A. Business Process Fragments Behavioral Merge[M]// On the Move to Meaningful Internet Systems: OTM 2014 Conference. Berlin: Springer, 2014: 112-129.

[21] 蒋宗礼，姜守旭. 形式语言与自动机理论[M]. 北京：清华大学出版社，2003：71-73.