

workflow网频繁子网挖掘研究进展^①

张书涵^{②*} 费超群^{*} 黄锡昆^{**} 李阳阳^{**}

(^{*} 中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)

(^{**} 中国科学院数学与系统科学研究院管理、决策与信息系统重点实验室 北京 100190)

(^{***} 中国科学院大学 北京 100049)

摘 要 本文总结了 workflow网频繁子网挖掘的主流研究方向,包括从一维的日志进程中构造 workflow网及其子网和从二维 workflow网中挖掘其频繁子网结构,总结了其中的代表性方法及其优缺点。 workflow网具有复杂、异构拓扑结构和完备性语义的特性,本文详细分析了将频繁模式挖掘(FPM)算法直接用于 workflow网频繁子网挖掘存在的问题及缺陷。并给出了 workflow网频繁子网挖掘的典型应用,包括异常检测、跨组织变体分析等。最后讨论了 workflow网频繁子网挖掘中的研究难点和未来研究趋势。

关键词 频繁模式挖掘(FPM); workflow网; 子网挖掘; Petri 网; 进程挖掘

0 引 言

workflow网频繁子网挖掘是指以 workflow网为输入,分析 workflow网中频繁出现的子结构(子模式),发现数据的内在规律,使其能够为工业上了解、优化、应用数据提供重要信息的一类问题^[1]。随着网络技术飞速发展, workflow网已经成为刻画业务执行逻辑的主要形式^[2],被广泛用于网络交易系统、业务流程管理等应用场景的建模和分析中^[3]。

当前 workflow网频繁子网挖掘研究在新型应用场景下面临着诸多挑战。例如,在业务流程管理领域,存在大量重复出现、高度可重用的子流程^[4],尤其是近年来,在跨组织 workflow网挖掘的场景下存在许多子网变体^[5]。变体的存在使得 workflow网频繁子网变得多种多样、纷繁复杂,缺乏通用性。

按照输入数据类型进行划分, workflow网频繁子网挖掘相关研究分为两大主流方向。其一,从一维的日志进程序列中实施进程挖掘,构造出 workflow网

及其子网^[6-7]。如何更高效地从日志中抽取、发现 workflow网及其子网,是当前研究的热点问题^[8-13]。另外,从二维的 workflow网中挖掘其频繁的子网结构^[1,14]。相关工作大多采用频繁模式挖掘(frequent pattern mining, FPM)相关方法进行求解。由于 workflow网中大量存在的选择、循环、并发等复杂模式及其特殊的语义约束,在 workflow网上进行频繁子网挖掘比传统的 FPM 问题更加复杂^[15]。

本文基于频繁模式挖掘的相关方法,总结了目前 workflow网频繁子网挖掘所取得的研究成果。详细分析了将传统 FPM 用于 workflow网频繁子网挖掘存在的问题及挑战,分析了基于 workflow网进行频繁子网挖掘与基于一般图进行频繁子图挖掘的本质区别。本文的主要贡献包括以下几个方面。

(1) 针对输入是一维日志序列的情形,研究方法分为基于过程式模式检测、基于日志转换和基于声明式规则检测 3 类方法,本文给出了各代表性方法,总结了优缺点。

(2) 针对输入输出均是二维 workflow网的情形,

① 国家重点研发计划(2016YFB1000902),国家自然科学基金(61232015,21472412,61621003)和中国博士后科学基金(2020TQ0341)资助项目。

② 女,1991 年生,博士生;研究方向:数据挖掘,Petri 网挖掘,机器学习;联系人,E-mail: zhangshuhan@ict.ac.cn。
(收稿日期:2021-03-17)

研究方法分为基于工作流网转换、基于序列与网混合式两类方法,本文给出了各代表性方法,总结了优缺点。

(3) 从理论上分析了传统频繁模式挖掘方法在工作流网频繁子网挖掘问题中的不适用情况及其根本原因,总结了工作流网频繁子网挖掘与频繁子图挖掘的本质区别。

本文剩余部分组织结构如下。第 1 节简要介绍相关的频繁模式挖掘方法;第 2 节给出工作流网频繁子网挖掘相关研究的两大主流方向,包括基于一维日志进程构造工作流网及其子网的研究方法和基于工作流网的频繁子网挖掘研究方法;第 3 节分析将传统 FPM 用于工作流网频繁子网挖掘存在的问

题及挑战;第 4 节列举分析了工作流网频繁子网挖掘中的几种典型应用;第 5 节总结了工作流网频繁子网挖掘技术的发展趋势及存在的挑战。

1 频繁模式挖掘

为了具体阐述工作流网频繁子网挖掘相关研究方向的分类,本节针对其使用的传统频繁模式挖掘方法进行梳理,详见图 1。频繁模式挖掘是数据挖掘中重要的研究问题,从最初的分析顾客购买记录到近几年用于进程挖掘,FPM 一直有着十分广泛的应用^[8,10,16-19]。依据不同的数据类型,相关方法分为频繁项挖掘、序列模式挖掘和频繁子图挖掘。

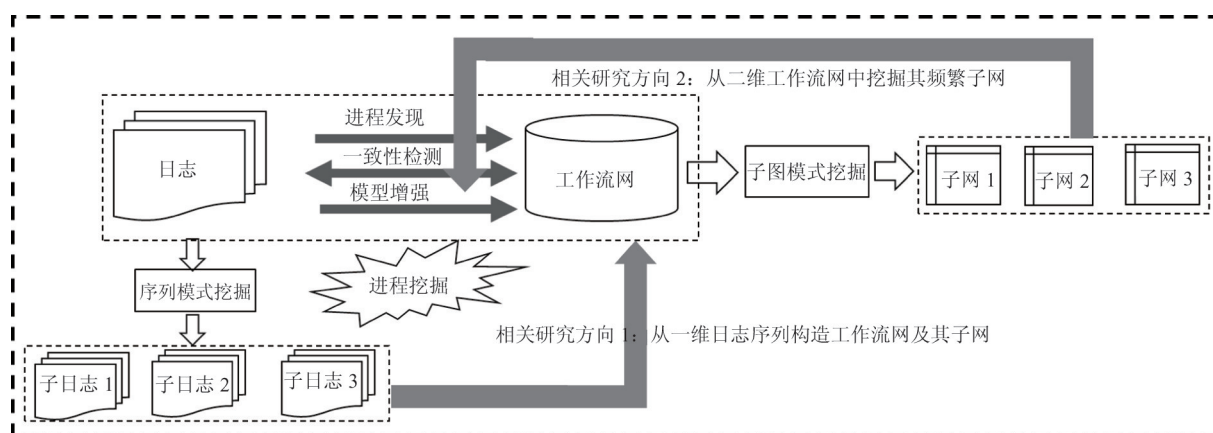


图 1 工作流网频繁子网挖掘研究方向分类

1.1 频繁项挖掘

频繁项挖掘(frequent itemset mining, FIM)问题提出之初旨在分析购物车中经常被一起购买的物品,得到顾客的购买习惯,为商店提供决策服务^[20]。FIM 问题的首个解决算法是 Apriori 算法^[21],其核心在于向下反单调性的保障,即如果一个项集不频繁,那么它的超集必定不频繁;反之,如果一个项集是频繁的,那么它的子集必定是频繁的。该性质可用偏序 $\leq$ 表示,即给定任意模式 p_1 及 p_2 , 如果 $p_1 \leq p_2$ 且 p_2 为频繁的,则 p_1 也是频繁的。向下反单调性是指导 FPM 及工作流网频繁子网挖掘的关键理论支撑。利用该性质,可以大大减小搜索空间。作为 FPM 最经典的算法框架,Apriori 具有良好的可扩展性。近年来,高性能计算、硬件加速等新技术被融合至

Apriori 算法。目前,已有数百个后续改进工作^[22-28]。

1.2 序列模式挖掘

序列数据库由有序元素或事件组成,典型的序列数据包括顾客购物序列、DNA 序列、时间序列等^[16,29-36]。随着数据类型的多样化,序列模式挖掘(sequential pattern mining, SPM)算法处理的对象也越来越复杂。首个处理 SPM 问题的算法为广义序列模式(generalized sequential pattern, GSP)挖掘算法^[16]。GSP 遵循逐层遍历的方式,但候选集的生成十分耗时。参考 FP-Growth 算法^[37]提出的模式增长计算方式,Jian 等人^[38]提出了 PrefixSpan 算法,以深度优先遍历的方法计算候选子序列的支持度。此外,PrefixSpan 利用向下反单调性提出了一种有效的

剪枝策略,减小搜索空间。随着序列数据维度及类型的复杂化,许多约束被融合到 SPM 问题的解决方法中。例如, Pex-Spam 算法使用的时间间隔约束^[39], MaxSP 算法使用的长度约束^[40]等。

1.3 频繁子图挖掘

作为数据结构中最为复杂的一种,图结构广泛应用于生物数据分析、社交网络分析等实际问题的数据建模中^[41-45]。Cook 和 Holder^[46]提出频繁子图挖掘(frequent subgraph mining, FSM)问题,旨在从给定的图集中抽取出频次大于一定阈值的子图。

首个解决 FSM 问题的算法是基于 Apriori 的图挖掘(Apriori-based graph mining, AGM)算法^[47],采取逐层两两连接的方式。如图 2(a)所示,大小为 k 的所有频繁子图连接生成大小为 $(k+1)$ 的候选子图,利用向下反单调性删除候选子图中的非频繁子图。AGM 类方法代表性的工作包括 AGM/AcGM 算法^[47]、FSG 算法^[48]、Path-Join 算法^[49]等,但不足之处是生成大量的候选集,需要耗费大量的内存。

基于以上不足,基于模式扩充的方法特点是在 k 级频繁子图 p 的基础上,每次扩展一条边,生成 p 的超图模式,然后计算支持度得到 $(k+1)$ 级频繁子图。从每个 $(k+1)$ 级频繁子图开始继续扩展,直到无法再扩展为止。此方法是一种全内存的计算,工作机制如图 2(b)所示,代表工作包括 gSpan 算法^[44]、GASTON 算法^[50]、SPIN 算法^[51]等。

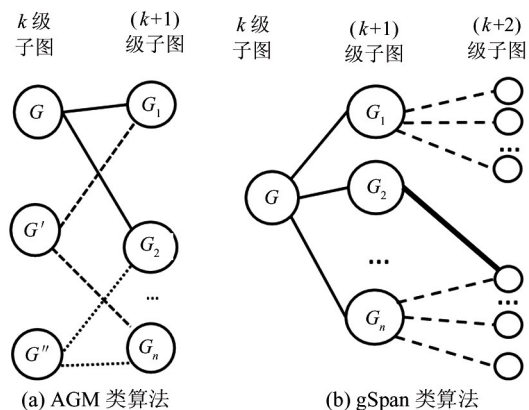


图2 频繁子图挖掘两种主要方法

2 工作流网频繁子网挖掘

按照输入数据类型进行划分,工作流网频繁子

网挖掘相关研究分为两大主流方向。其一,从一维的日志进程序列中实施进程挖掘,构造出工作流网及其子网^[6,7]。另外,从二维的工作流网中挖掘其频繁的子网结构^[1,14]。

2.1 基于一维日志进程构造工作流网及其子网

进程挖掘在文献[52]中被提出,IBM 艾曼登研究中心将进程发现思想引入业务过程管理中^[53]。进程挖掘旨在以大规模日志为原始输入,从中抽取系统运行过程,通常以工作流网的形式呈现。随着业务流程管理系统的兴起,日志信息日益丰富,进程挖掘方法的重要性逐步提升^[6,7,54-56]。当前,进程挖掘任务分为 3 种类型,即进程发现、一致性检测和模型增强,如图 1 所示。而每种任务要解决的核心问题是确保模型能够“重演”。进程发现问题主要采用基于规则匹配的方法^[55]进行求解。一致性检测问题主要采用分解日志的方法进行对齐计算^[57]。模型增强问题关注日志中时间、资源等组织信息的提取。

由于进程挖掘和序列模式挖掘具备问题的相似性,已有许多 SPM 方法被用于进程挖掘中^[8-11,58-62]。

目前,日志中包含的关系可分为两类:

(1)控制流依赖关系,即日志中包含的所有活动都遵循一定的次序,例如直接依赖、选择、并发关系。

(2)声明式约束,即日志中的活动可用时态逻辑描述,如发生的情况最终一定会发生。

根据日志关系类型划分,基于进程挖掘构造工作流网及其子网的研究方法可分为基于过程式模型检测、基于日志转换和基于声明式检测的方法。

2.1.1 基于过程式模型检测的方法

定义 1 过程式模型检测问题^[55]

给定一组活动集合 A , 由一系列活动组成的非空序列称为工作流执行轨迹 $\sigma \in A^*$, 轨迹的多重集合表示为日志 $L \in B(A^*)$, 其中 $B(A^*)$ 表示 A^* 的幂集。过程式模型是指事件次序与日志中的活动次序保持一致,如顺序模式、并行分支模式、同步模式等。过程式模型检测问题是指将日志转化为端到端过程模式的过程。

表1列举了过程式模型检测任务清单,各任务

表 1 过程式模型检测问题任务清单

具体任务	使用方法	算法	工具	衡量指标
工作流网构造	直接后继关系抽取	系列算法 ^[55]	——	——
工作流网预测	频繁日志转换为 workflow 图	w-find, c-find ^[63]	——	#Candidates Runtime
工作流网抽象	频繁日志抽取作预处理	Preprocessing ^[64]	HeuristicMiner	Fitness
异常模式检测	非频繁日志抽取作预处理	iInductiveMiner ^[65]	InductiveMiner	Rediscoverability
片段发现	频繁日志转换为 DAG	EpisodeDiscovery ^[9]	EpisodeDiscovery	#Episodes #Candidates Runtime
子结构发现	频繁日志转换为实例图	IGP ^[59]	ProM	#Top subs, COV, DIV Support Confidence
局部信息获取	频繁日志转换为过程树	LPM ^[58]	LocalProcess ModelDiscovery	Language Fit Determinism Coverage
统计模式检测	频繁日志随机取样聚类	SubgroupDiscovery ^[66]	SubgroupDiscovery	Case study Correlation
目标驱动检测	高可用函数剪枝过程树构造	HU-LPM ^[67-68]	LocalProcess ModelDiscovery	Case study #Activities Utility function #Runtime
相关性模型检测	带间隔约束的频繁日志抽取	Constraint-basedLPM ^[69]	LocalProcess ModelDiscovery	Runtime

均是以序列模式挖掘方法为解决方案。其中,最具代表性的任务为频繁子片段 (episode) 挖掘问题。

定义 2 片段,子片段^[9]

片段为事件的偏序集合,表示为三元组 $\alpha = \langle V, \leq, g \rangle$, 其中 V 为一组事件的集合, $\leq$ 为定义在 V 上的偏序关系集合, $g: V \rightarrow A$ 为事件和活动之间的对应函数。存在另一个片段 $\beta = \langle V', \leq', g' \rangle$ 是 α 的子片段, 记作 $\beta \leq \alpha$, 当且仅当存在内射函数 $f: V' \rightarrow V$, 使得以下 2 个条件同时满足:

- (1) $\forall v \in V': g'(v) = g(f(v))$, 即 β 中的所有节点都出现在 α 中;
- (2) $\forall v, w \in V' \wedge v \leq' w: f(v) \leq f(w)$, 即 β 中的所有边都出现在 α 中。

片段使用有向无环图 (directed acyclic graph, DAG) 进行表示, DAG 中的节点代表事件 (出现在日志的活动集合中), DAG 中的边代表事件之间存在偏序关系。如图 3 所示为 3 个片段, 其中 E_1 表示事

件 A(D) 与事件 B(E) 并发执行, 事件 C 只能在事件 A 和 B 发生之后执行; E_2 表示事件 A 执行后有事件 B 和事件 C 两种选择, 而事件 B 执行后 D 才能执行, 或者事件 C 执行后 D 才能执行; E_3 表示, 事件 A 和 B 并发执行, 其中 A 和 B 之间没有连边。

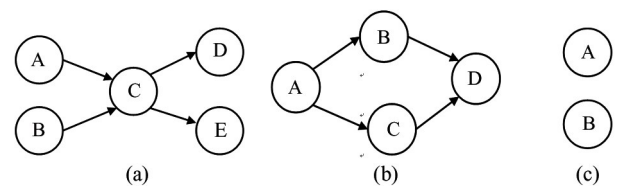


图 3 片段示例

定义 3 频繁子片段挖掘问题^[9]

片段 $\alpha = \langle V, \leq, g \rangle$ 在工作流执行轨迹 $\sigma = \langle a_1, a_2, \dots, a_n \rangle$ 中出现, 记作 $\alpha \subseteq \sigma$, 当且仅当存在一个内射函数 $h: V \rightarrow \{1, \dots, n\}$, 使得以下 2 个条件同时满足:

(1) $\forall v \in V: g(v) = a_{h(v) \in \sigma}$ 成立, 即片段中的事件与日志中的活动一致;

(2) $\forall v, w \in V \wedge v \leq w: h(v) \leq h(w)$, 即片段中事件之间的偏序关系与日志中活动间的次序一致。

片段 α 在日志 $L \in B(A^*)$ 中出现的频率定义为

$$freq(\alpha) = \frac{|\{\alpha \in L \mid \alpha \subseteq \sigma\}|}{|L|}$$

表示片段中的偏序关系在日志中出现的次数。

片段发现问题旨在从事件日志中找到所有满足 $freq(\alpha)$ 大于等于给定阈值的片段。

2.1.2 基于日志转换的方法

求解过程式模型检测问题, 通常采用基于转换的方法, 即将频繁日志转换为图/树模式。(1) 工作流图是最早提出的频繁日志的转换对象^[53], 它的特点是不能识别工作流中的 AND/OR 汇集结构和分支结构。因此, 工作流图只能用来表示不包含 OR 结构的工作流网。(2) 有向无环图 (directed acyclic graphs, DAG) 是最常用的转换表示^[9]。然而目前, DAG 只能表示频繁日志中的顺序和并发模式, 对于日志中出现的选择、循环等复杂结构仍不能检测。(3) 实例图常用于聚类, 实施步骤为首先将每条频繁日志转换为实例图, 然后运行图聚类算法^[59]。与 DAG 类似, 实例图也仅能表示频繁日志中的顺序和并发模式。(4) 过程树是一种直观的树型结构, 树中的叶子节点代表日志中的活动, 分支节点代表活动之间的关系^[70]。使用过程树构造 Petri 网的步骤是, 首先从频繁日志中提取过程树, 然后根据 6 种预设规则, 将过程树转换为可靠的 (也称为良构的) 端到端过程模型。与前几种表示不同, 基于过程树的方法不仅支持基本的控制流模式, 还可表示复杂的选择、循环等结构。

2.1.3 基于声明式规则检测的方法

近年来, 由进程挖掘中的进程发现任务衍生出了另一项具有挑战性的任务——声明式规则发现 (declarative rule discovery)^[71], 是指从日志中提取一系列符合时态逻辑规则的任务, 又称约束发现。已往的进程挖掘中的进程发现任务, 只能得到结构化 (structured) 的工作流网, 而声明式规则是一种弱结构化 (less-structured) 的表示形式, 是对进程挖掘

任务的补充。

定义 4 声明式规则发现^[72]

给定活动集合 $A = \{a_1, a_2, \dots, a_m\}$ 和一条日志 $L = \{t_1, t_2, \dots, t_n\}$, 其中 $t_q (1 \leq q \leq n)$ 表示一条取值于 A 的有限活动记录。声明式规则挖掘旨在从日志中抽取满足时态逻辑关系的一系列约束。其中, 约束由以下 3 部分组成。

(1) 取值于 A 的活动。

(2) 活动之间的布尔型连接关系, 包括非 ($\neg$)、与 ($\wedge$)、或 ($\vee$)、蕴含 ($\Rightarrow$)、等价 ($\Leftrightarrow$)。

(3) 基于有限序列的线性时态逻辑 (linear temporal logic on finite traces, LTL_f) 算子, 包括必然算子 ($\square$)、可能算子 ($\diamond$)、直到算子 ($\cup$)、下一时刻算子 ($\circ$)。

例 1 给定活动集合 $A = \{S, A\}$ 和日志 $t = \{SSSASASSA, 123456789\}$, 其中 $1, 2, \dots$ 代表活动在日志中的下标。日志 t 中包含一条声明式规则 $RESPONSE(S, A)$, 其含义为 $\square(S \Rightarrow \diamond(A))$, S 发生时都有 A 跟随发生。

日志 t 中规则 $RESPONSE(S, A)$ 的实例为 $\{S_1A_1\}, \{S_2A_1\}, \{S_3A_1\}, \{S_4A_2\}, \{S_5A_3\}, \{S_6A_3\}$ 。

在声明式规则发现任务中, 仍然是以日志作为主要输入数据类型。但不同的是, 关注日志的视角和粒度与传统进程挖掘不同。日志中记录着大量已经发生的活动 (也称事件、案例、任务) 以及活动间的逻辑依赖关系。这些依赖关系是联系日志与进程模型、日志与 Declare 规则之间的重要纽带^[73]。作为常用的数据预处理方法, FIM 方法首先针对日志中大量的活动进行初步过滤, 然后提取活动之间的依赖关系。FIM 问题与声明式挖掘问题的数据类型有着密不可分的联系, 如表 2 所示。表 3 列举了声明式规则检测问题任务清单。

表 2 频繁项挖掘与声明式规则挖掘对比

对比指标	频繁项挖掘	声明式规则挖掘
输入数据	事务	日志/记录
处理对象	项/项集	活动/活动集
关系描述	偏序	控制流依赖
输出对象	关联规则	隐含约束

表 3 声明式规则检测问题任务清单

具体任务	使用方法	算法	工具	衡量指标
行为匹配	频繁日志转换为实例图	GPD ^[74]	RapidProM	Case study runtime
约束抽取	日志聚类,关联规则挖掘	Data perspective DPM ^[71]	Reproducibility Siren	#Rules, Support Confidence

2.2 基于二维 workflow 网挖掘频繁子网

定义 5 频繁 workflow 网子网挖掘

给定 workflow 网集 $W = \{w_1, w_2, \dots, w_T\}$, 子网 x 的支持度定义为包含它的工作流网占比

$$\text{sup}(x, W) = \frac{|\{W_i \mid x \subseteq w_i \wedge w_i \in W\}|}{T}$$

频繁 workflow 子网挖掘旨在从 workflow 网集中挖掘得到子网/片段,以供 workflow 网重用、设计。workflow 网频繁子网挖掘的相关研究重点关注以下两方面问题。

(1) workflow 网标签一致性。workflow 网的标签往往使用 WordNet 进行统一规范,标签有固定的语法形式,即动词 + 宾语,例如检查订单、创建账户等。

(2) workflow 网适合的图形表示方法为基于转换的方法。本文主要关注此类问题。

2.2.1 基于 workflow 网转换的方法

目前已有的 workflow 网频繁子网挖掘通常采用两步走的解决方案。

第一步,对 workflow 网进行转换。例如,(1)仅将 workflow 网中的事件转换为图节点,将条件以及输入输出弧作为图的边,多用于转换一种特殊类型的工作流网—标识图。(2)仅将 workflow 网中的条件转换为图节点,将事件、输入输出弧作为图的边^[75],多用于另一类 workflow 网—状态机。(3)将 workflow 网中的两类节点(事件、条件)转换为图的一种节点,将输

入输出弧作为图的边^[76]。

第二步,基于转换的图直接运行频繁子图挖掘、图聚类算法。首个工作流网挖掘方法是 w-find 算法^[77]。该算法的输入是 workflow 转换而成的控制流图(control flow graph, CFG),算法的目标是挖掘出 CFG 中所有满足闭合约束的子图。然而,CFG 只包含一类节点,代表日志中的活动,边代表活动之间的数据流。可以发现,w-find 算法利用频繁子图挖掘技术得到频繁的活动,是对进程挖掘任务的验证和扩充。文献[78]提出 BPMClustering 算法,利用图聚类方法基于多个 workflow 网进行聚类计算,计算多个 workflow 网之间的相似度。然而,该方法不能挖掘常见的选择和并发结构。文献[79]提出 IGP 算法,该算法将每条日志都转换为一个实例图(instance graph),然后在大规模的实例图之间实施频繁子图挖掘,但挖掘出的实例图仅能还原顺序和并发结构,一些复杂模式无法挖掘。类似 w-find 算法,IGP 算法的目标也是对于进程挖掘任务进行部分验证。文献[1]基于 w-find 算法提出从进程模型中挖掘频繁模式的问题,将频繁子图挖掘的方法应用于 workflow 网模式抽取中,并开发出工具 WoMine 对 workflow 网的拓扑结构进行挖掘。然而算法得到的子结构不再是 workflow 网。常见的基于图集合的频繁子图挖掘算法在 workflow 网挖掘中的应用如表 4 所示。

表 4 基于图集合的频繁子图挖掘算法在 workflow 网挖掘中的应用

算法	关键技术	顺序	并行	选择	循环	间接依赖	隐形模式	重名模式
w-find ^[77]	挖掘闭合控制流图	+	+	-	-	-	-	-
BPMClustering ^[78]	图聚类	+	-	-	-	-	-	-
IGP ^[79]	基于实例图的聚类	+	+	-	-	-	-	-
WoMine ^[1]	w-find 算法	+	+	+	+	+	-	-

2.2.2 序列与网混合式方法

混合的方法是指从一维日志序列中构造二维的

workflow 网,然后再将 workflow 网与日志序列进行对比计算。代表性的混合方法是一致性检测算法^[80]。

一致性检测利用对齐映射等方式解决模型重演问题,但是大多数一致性检测技术仅提供基本统计信息,例如不一致的偏差数量。关于序列与网混合的

方法研究仍是少数,表 5 中列举了几种代表性的序列与网混合方法。

表 5 序列与网混合式方法所适用的具体任务清单

具体任务	使用方法	算法	工具	衡量指标
过程式一致性检测	结合拓扑结构与序列中的行为特征进行一致性检测	Recursive invisible Duplicated ^[81]	Conformance Checker in ProM	Fitness Appropriateness
大规模一致性检测	利用分治策略分别将日志进行划分,将进程模型进行划分	RPST ^[82]	PLG in ProM	Fitness
基于熵的一致性检测	使用熵重新定义了子模式的准确率、召回率等指标	Entrapia ^[83]	Entrapia	Precision Recall
声明式一致性检测	基于整数线性规划的策略求解最优的比对方案	MPCC ^[57]	ProM	#Constraints Fitness, runtime

业务流程的执行顺序决定了业务的正确性和完整性。检测此工作流网与日志序列之间的一致性程度,关乎业务流程的正常运行。但如果单独的检测每个流程的偏差,十分耗时耗力。利用频繁子网的方式将工作流网中重复执行的片段进行提取,利用分治的策略将日志进行分解,是一种常用的技术手段^[82]。

例如,银行的汇款流程预定义了检测步骤:开始汇款处理、读取系统配置文件、核查收款人的银行信息等,完整的流程在每次汇款时都会重复发生。

实际上,在上述重复执行的汇款进程中通常存在一种异常模式,即具有高级权限的 VIP 客户可以跳过冗长的预配置检测阶段。然而这种异常模型并非每次都出现在日志序列中,这就需要使用频繁模式挖掘的手段对工作流网的子结构进行混合分析。

3 FPM 用于工作流网频繁子网挖掘的问题及缺陷分析

3.1 基于 SPM 构造工作流网的问题

在进程挖掘研究中,关注更多的是如何从日志中抽取满足偏序关系、控制流依赖关系的工作流网。Leemans 和 Aalst^[9]提出 EpisodeDiscovery 算法,利用序列挖掘算法 PrefixSpan 分析日志序列。但要求所采用的日志必须具备时间、资源等具有偏序关系的数据。EpisodeDiscovery 算法的目的是从日志中抽

取出工作流网的顺序结构,体现出与日志中一致的偏序关系。Tax 等人^[58]提出 LocalProcessModel 算法,旨在从一维的日志序列中构造出工作流网的顺序结构、并发结构、选择结构等复杂的结构。该算法的目的与前面的方法均有所不同,算法并不要求得到完整的工作流网,只需要得到局部结构作为重要的模式。

尽管当前序列模式挖掘、进程挖掘问题已被广泛研究,但将序列模式挖掘方法用在进程挖掘,解决工作流网及其子网的构造仍存在以下问题。

(1)构造的工作流网结构复杂。由于进程挖掘的诸多方法基于规则匹配,存在封闭世界的假设,使得构造出的许多工作流网存在过拟合的现象,存在许多结构复杂、重复、冗长的工作流网模型。在大数据的场景下,文献[84]中提出利用分治策略将日志和模型进行分解,从而降低进程挖掘构造工作流网的复杂度,但挖掘到的工作流网仍然十分复杂,许多研究者称之为“意大利面”式。

(2)声明式规则发现低效。由于声明式规则使用基于有限序列的线性时态逻辑 LTL_f 的操作算子进行表示,当规则数量庞大时,算法处理十分耗时。但声明式规则的优点是比较灵活,适合于不同层面的日志解析,不再局限于以结构为导向。

(3)可解释性差。许多挖掘出的频繁子结构/子模式的意义难以解释,因为缺乏领域知识的支撑,通常被认作无意义。例如,在生物信息学中,挖掘出

的频繁模式是否有意义需要由领域专家决定。

3.2 基于 FSM 进行 workflow 网挖掘的问题

workflow 网是一种具有单一入口、单一出口,并且执行流程受结构化约束的 Petri 网,是刻画业务执行逻辑的主要形式。Petri 网中的元素“条件”、“事件”、“流关系”与 workflow 网中的“案例”、“任务”、“过程”一一对应。

定义 6 workflow 网^[55]

令 $N = (C, E; F)$ 一个 Petri 网, N 是 workflow 网, 若

(1) $\forall e \in E: \cdot e \neq \emptyset \wedge e \cdot \neq \emptyset$;

(2) $\exists i \in C: |\{i | i \cdot = \emptyset\}| = 1$, 其中 i 称为 N 的源条件;

(3) $\exists o \in C: |\{o | o \cdot = \emptyset\}| = 1$, 其中 o 称为 N 的汇结条件;

(4) $\forall x \in C \cup E$ 都位于从 i 到 o 一条正向路径或反向路径上。

由于频繁子图挖掘算法在一般图上的成功应用,许多研究者致力于将 FSM 技术推广到其他领域。在 workflow 网上的应用需要将 workflow 网进行转换。然而大多数情况下,这种转换无法还原,不满足双射关系,会丢失重要信息。因此,传统的频繁子图挖掘方法无法直接用于 workflow 网挖掘。其本质原因如下。

(1) 复杂、异构拓扑结构。从定义 6 中不难看出,workflow 网具有两类节点集和一类边集,而一般图通常由一类节点集和一类边集组成。通常情况下,workflow 网的第一类节点叫做事件(表示动作)节点,第二类节点叫做条件(表示状态)节点。

(2) 非独立带约束连接类型。遵循 Petri 网的结构性质,在 workflow 网中,弧(连接)是不能独立的,必需依附在它两端的条件/事件节点,且同一类型的节点之间不能有弧。这也使得 workflow 网的结构约束性更强,是 workflow 网处理中高复杂度的主要来源。

(3) 完备性语义。一般图结构可无语义约束(或可赋予任意语义)。Petri 教授本人曾指出, Petri 网具有内置完备性语义。一个 Petri 网中的流关系通常是“前提条件事件后提条件”。此内置语义为 Petri 网中不可分割的单元。然而转换为一般图后

丢失了完备性语义。workflow 网是 Petri 网的一种特殊子类,因此,workflow 子网应是一种不可分割的执行单元^[15]。

(4) 执行约束。定义 6 的 2、3 条件规定了 workflow 网具有唯一的入口和出口,即 workflow 网具有执行约束^[5],此约束是 workflow 网满足可靠性、无死锁的重要理论依据。

目前,workflow 网挖掘工作得到的子网都不再是 workflow 网,因为转换破坏了 workflow 网的语义特性。从理论角度分析,目前缺乏关于 workflow 子网的统一定义,通常将其与一般图等同处理。进一步讲,缺乏 workflow 子网同构的严格定义。这些都是 workflow 频繁子网挖掘工作中有待探索的问题。

4 workflow 网频繁子网挖掘典型应用

4.1 异常检测

对于构建 workflow 网及挖掘 workflow 网结构这两类主要任务而言,异常模式的存在都为问题求解增加了难度。许多异常体现出统计意义上的特征,例如出现的较少(低于频繁阈值),即可被判定为异常。已有很多研究成果利用 FPM 方法对异常进行检测、过滤处理^[85-87]。基于关系对日志过滤、基于统计概率对日志过滤,都在一定程度上提升了日志质量,使得构造的 workflow 网质量有所保障。

4.2 跨组织变体分析

在业务流程管理(business process management, BPM)领域,广泛存在子结构重复出现的现象。它们具有高度重用的特点。跨组织变体问题是指针对同一个流程,在不同组织、不同地域之间会存在着共性和差异性,这些差异性被称为变体^[6]。变体中的共性信息体现了不同变体之间的关联,得到共性信息可以促进不同组织之间的信息共享、甚至是平台共用。例如,对于医生诊断流程,不同医院和不同医生检查在诊断步骤上会存在一些差别,但不同的医院、不同医生之间往往遵循一个公共的模式^[88]。得到这种公共模式可以为许多医院提供病例模版和诊断指导。

4.3 模型重用

对于大多数制造企业来说,设计新的业务流程

(如零件装配流水线) 耗费巨大, 需要多部门协调、整合历史资源。文献[89]针对不同组织环境、不同部门之间相同/相似的子结构/子模式进行相似度计算, 同时也表明了已有 Petri 网模式的重用对于企业的流程设计而言意义重大。生物信息学领域提出 workflow 数据集 myExperiment^[90]、Taverna^[91] 及其配套的开发软件, 旨在检索、共享、重用 workflow。利用频繁模式挖掘 (FPM) 方法提取频繁的子结构是常用的研究方法。

4.4 其他应用

除上述应用外, 频繁模式挖掘技术也在不断进步, 提出了更能够体现模式意义的算法^[67, 92-94]。频繁模式挖掘技术在不同领域中也发挥着重要作用, 例如医疗辅助保健系统、决策系统以及 e-learning 电子学习系统。在这些系统中, 软件执行过程通常存在重复、循环^[95-96]。在 e-learning 系统中, 通过学生与课堂之间的互动分析频繁的行为模式^[1]。另一方面, 老师可以根据挖掘的频繁模式改进课程设计。

5 总结与展望

本文首先依据 3 种数据类型, 对于 workflow 网频繁子网挖掘的研究方法进行分类, 其次, 针对 2 种主流研究方向, 即从一维日志进程中构造 workflow 网及其子网和从二维 workflow 网中挖掘其频繁子网结构, 分别给出具体的问题描述、任务列表及相关数据表示。本研究可得到如下结论。

(1) 在以进程挖掘技术作为手段, 构建 workflow 网的任务中, 诸多 SPM 算法被应用, 聚焦于 workflow 网在控制流层面 (简单的拓扑结构) 的构造。然而针对复杂拓扑结构, 如选择、循环等结构以及不可见的隐含信息, 当前 SPM 算法仍然不能平衡准确度和泛化程度, 并且容易过拟合。

(2) 在进程挖掘技术的演进发展中, 构建声明式的规则为一项比较新颖的任务。FIM 算法被用于日志预处理阶段, 其关键在于保持日志活动序。然而, 目前只有少量 FIM 算法被用于此任务中。

(3) 对于具有二维异构拓扑结构的 workflow 网挖掘分析问题, 解决方法更为复杂。当前, 经典的

FSM 算法, 如 gSpan 算法、SUBDUE 算法被广泛用于挖掘 workflow 网子网问题中。然而, 由于缺乏针对 workflow 网的理论问题探究, 例如, 对于 workflow 网结构的特殊性、workflow 网表示方法的正确性等的考察, 适合于 workflow 网的频繁子网算法还有待严格的理论验证。

伴随着网络技术的飞速发展, 网络交易系统、业务流程管理系统等应用场景产生了海量大规模 workflow 网。复杂异构的网络结构、丰富的语义, 为 FPM 在 workflow 网中的应用带来了新的挑战。将 workflow 网频繁子网挖掘研究推广至一般 Petri 网更具难度。因为一般 Petri 网类型众多, 性质更复杂。文献[15]提出完备子网的概念和适用于一般 Petri 网的压缩数据表示方法, 是对一般类型 Petri 网的频繁子网挖掘研究的尝试。但是文献[15, 97]中的一般 Petri 网频繁子网挖掘技术仍不能用于 workflow 网, 这是因为缺乏保持特殊语义的划分 workflow 子网的方法。因此, 以下几点是 workflow 网频繁子网挖掘研究可能的发展趋势:

(1) 针对海量数据特征, 设计一种包容异构性、完备性语义的 workflow 网子网划分方法, 以及满足完备性语义的新的表示形式, 并设计更高效的工作流网挖掘算法, 将频繁模式挖掘理论推广至 workflow 网及其广义领域。

(2) 一直以来, 在 Petri 网 (包含 workflow 网) 研究领域中, 带 token (托肯, 也被称为令牌) 的动态信息一直是 workflow 网性能分析的瓶颈^[98]。特别是, 托肯体现全局范围内资源的流动, 而子网技术是局部层面的研究。文献[98]考虑动态性质的同时探讨了网络交易系统漏洞检测问题的解决思路。如何设计子网挖掘算法, 保持整个 workflow 网的性质, 并在电信诈骗检测、用户行为分析等更具挑战性的场景中发挥作用, 是未来工作的重点。

(3) workflow 网是一种特殊类型的 Petri 网, 而 Petri 网区别于其他建模工具的根本性优势在于其能够描述真并发、具有严格的执行语义。除了本文聚焦的拓扑层面的完备性语义, Petri 网通用网论 (包括多种 Petri 网的共有特性挖掘, 如并发特性、同步特性、冲突特性) 还有待进一步研究。

参考文献

- [1] CHAPELA-CAMPA D, MUCIENTES M, LAMA M. Mining frequent patterns in process models[J]. *Information Sciences*, 2019, 472: 235-257
- [2] AALST W M P. Structural characterizations of sound workflow nets[J]. *Computing Science Reports*, 1996, 9623: 22
- [3] PINNA A, TONELLI R, ORRÚ M, et al. A Petri nets model for blockchain analysis[J]. *The Computer Journal*, 2018, 61(9): 1374-1388
- [4] AALST W M P, TER HOFSTEDE A H M, WESKE M. Business process management: a survey[C] // Proceedings of the 1st International Conference on Business Process Management, Eindhoven, Netherlands, 2003: 1-12
- [5] LIU C, ZENG Q, CHENG L, et al. Privacy-preserving behavioral correctness verification of cross-organizational workflow with task synchronization patterns[J]. *IEEE Transactions on Automation Science and Engineering*, 2020, pp(99):1-12
- [6] AALST W M P, ADRIANSYAH A, MEDEIROS A K A, et al. Process mining manifesto[C] // International Conference on Business Process Management, Clermont-Ferrand, France, 2011: 169-194
- [7] 曾庆田. 过程挖掘的研究现状与问题综述[J]. *系统仿真学报*, 2007, 19(a01): 275-280
- [8] SANI M F, ZELST S J V, AALST W M P. Applying sequence mining for outlier detection in process mining[C] // OTM Confederated International Conference 'On the Move to Meaningful Internet Systems', Valletta, Malta, 2018: 98-116
- [9] LEEMANS M M, AALST W M P. Discovery of Frequent Episodes in Event Logs[M]. Vienna: Springer International Publishing, 2015: 1-31
- [10] GENGA L, ALIZADEH M, POTENA D, et al. Discovering anomalous frequent patterns from partially ordered event logs[J]. *Journal of Intelligent Information Systems*, 2018, 51(2): 257-300
- [11] TAX N, SIDOROVA N, AALST W M P, et al. Local process model discovery: bringing Petri nets to the pattern mining world[C] // International Conference on Application and Theory of Petri Nets and Concurrency, Bratislava, Slovakia, 2018: 374-384
- [12] HASSANI M, ZELST S J V, AALST W M P. On the application of sequential pattern mining primitives to process discovery: overview, outlook and opportunity identification[J]. *WIREs Data Mining and Knowledge Discovery*, 2019, 9(6): e1315
- [13] CHAPELA-CAMPA D, MUCIENTES M, LAMA M. Understanding complex process models by abstracting infrequent behavior[J]. *Future Generation Computer Systems*, 2020, 113: 428-440
- [14] DIAMANTINI C, POTENA D, STORTI E. Mining usage patterns from a repository of scientific workflows[C] // Proceedings of the 27th Annual ACM Symposium on Applied Computing, New York, USA, 2012: 152-157
- [15] LU R, ZHANG S. PSpan: mining frequent subnets of Petri nets[EB/OL]. <http://arxiv.org/pdf/2101.11972.pdf>; arXiv, (2021-01-28), [2021-03-01]
- [16] AGRAWAL R, SRIKANT R. Mining sequential patterns[C] // Proceedings of the 11th International Conference on Data Engineering, Taipei, China, 1995: 3-14
- [17] LI X, LI J, GAO H. An efficient frequent subgraph mining algorithm[J]. *Journal of Software*, 2007, 18(10): 2469-2480
- [18] COOK D J, HOLDER L B. Mining Graph Data[M]. New York: John Wiley & Sons, 2007: 99-109
- [19] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules in large databases[C] // Proceedings of the 20th International Conference on Very Large Data Bases, San Francisco, USA, 1994: 487-499
- [20] AGRAWAL R, Imieliński T, SWAMI A. Mining association rules between sets of items in large databases[C] // Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, New York, USA, 1993: 207-216
- [21] AGRAWAL R, SRIKANT R. Fast Algorithms for Mining Association Rules[M]. San Francisco: Morgan Kaufmann Publishers Inc., 1998: 580-592
- [22] BASHIR S. An efficient pattern growth approach for mining fault tolerant frequent itemsets[J]. *Expert Systems with Applications*, 2020, 143: 113046
- [23] CHEN Y, AN A. Approximate parallel high utility itemset mining[J]. *Big Data Research*, 2016, 6: 26-42
- [24] ZAKI M J, HSIAO C J. CHARM: An efficient algorithm for closed itemset mining[C] // Proceedings of the 2002

- SIAM International Conference on Data Mining, Chicago, USA, 2002: 457-473
- [25] ZAKI M J, HSIAO C J. Efficient algorithms for mining closed itemsets and their lattice structure [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2005, 17(4): 462-478
- [26] TSENG V S, WU C W, FOURNIER V P, et al. Efficient algorithms for mining top-k high utility itemsets [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2016, 28(1): 54-67
- [27] AMPHAWAN K, LENCA P, SURAREKS A. Efficient mining top-k regular-frequent itemset using compressed tidsets[C] // Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2011 International Workshops, Shenzhen, China, 2011: 124-135
- [28] FOURNIER V P, LIN J C W, DUONG Q H, et al. FHM + : faster high-utility itemset mining using length upper-bound reduction[C] // International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Trends in Applied Knowledge-Based Systems and Data Science, Morioka, Japan, 2016: 115-127
- [29] GARCÍA H R A, MARTÍNEZ T J. F J F, CARRASCO O J A. A new algorithm for fast discovery of maximal sequential patterns in a document collection[C] // International Conference on Computational Linguistics and Intelligent Text Processing, Mexico City, Mexico, 2006: 514-523
- [30] PEI J, HAN J W, WANG W. Constraint-based sequential pattern mining: the pattern-growth methods[J]. *Journal of Intelligent Information Systems*, 2007, 28(2): 133-160
- [31] YAN X F, HAN J W, AFSHAR R. CloSpan: mining: closed sequential patterns in large datasets[C] // Proceedings of the 2003 SIAM International Conference on Data Mining, San Francisco, USA, 2003: 166-177
- [32] ZHENG Z G, ZHAO Y C, ZUO Z Y, et al. Negative-GSP: an efficient method for mining negative sequential patterns[C] // Proceedings of the 8th Australasian Data Mining Conference, Darlinghurst, Australia, 2009: 63-67
- [33] ZIHAYAT M, CHENG W W, AN A, et al. Mining high utility sequential patterns from evolving data streams[C] // Proceedings of the ASE BigData & Social Informatics 2015, New York, USA, 2015: 1-6
- [34] ZIHAYAT M, DAVOUDI H, AN A. Mining significant high utility gene regulation sequential patterns[J]. *BMC Systems Biology*, 2017, 11(6): 109
- [35] ZHANG S H, HOU S L, FEI C Q. Cpriori: an index-based framework to extract the generalized center strings [C] // International Conference on Web Information Systems Engineering, Dubai, United Arab Emirates, 2018: 468-482
- [36] 张长海, 胡孔法, 陈峻, 等. 一种高效的基于位图序列模式挖掘算法[J]. *高技术通讯*, 2010, 20(2): 133-137
- [37] WANG J, HAN J W, LI C. Frequent closed sequence mining without candidate maintenance[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2007, 19(8): 1042-1056
- [38] JIAN P, HAN J W, MORTAZAVI A B, et al. PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth[C] // Proceedings of the 17th International Conference on Data Engineering, Washington, USA, 2001: 215-224
- [39] HO J, LUKOV L, CHAWLA S. Sequential pattern mining with constraints on large protein databases[C] // Proceedings of the 12th International Conference on Management of Data, Hyderabad, India, 2005: 89-100
- [40] FOURNIER V P, WU C W, TSENG V S. Mining maximal sequential patterns without candidate maintenance [C] // International Conference on Advanced Data Mining and Applications, Hangzhou, China, 2013: 169-180
- [41] JIANG C, COENEN F, ZITO M. A survey of frequent subgraph mining algorithms[J]. *The Knowledge Engineering Review*, 2013, 28(1): 75-105
- [42] KURAMOCHI M, KARYPIS G. An efficient algorithm for discovering frequent subgraphs[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2004, 16(9): 1038-1051
- [43] YAN X F, HAN J W. CloseGraph: mining closed frequent graph patterns[C] // Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, 2003: 286-295
- [44] YAN X F, HAN J W. gSpan: graph-based substructure pattern mining[C] // IEEE International Conference on Data Mining, Maebashi City, Japan, 2002: 721-724

- [45] 邹兆年, 李建中, 高宏, 等. 从不确定图中挖掘频繁子图模式[J]. 软件学报, 2009, 20(011): 2965-2976
- [46] COOK D J, HOLDER L B. Substructure discovery using minimum description length and background knowledge [J]. *Journal of Artificial Intelligence Research*, 1994, 1: 231-255
- [47] INOLUCHI A, WASHIO T, MOTODA H. An apriori-based algorithm for mining frequent substructures from graph data [C] // European Conference on Principles of Data Mining and Knowledge Discovery, Lyon, France, 2000: 13-23
- [48] KURAMOCHI M, KARYPIS G. Frequent subgraph discovery [C] // Proceedings of 2001 IEEE International Conference on Data Mining, San Jose, USA, 2001: 313-320
- [49] 刘勇. 图模式挖掘技术的研究[D]. 哈尔滨: 哈尔滨工业大学计算机科学与技术学院, 2010: 22-23
- [50] NIJSSEN S, KOK J N. A quickstart in frequent structure mining can make a difference [C] // Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, 2004: 647-652
- [51] HUAN J, WANG W, PRINS J, et al. Spin: mining maximal frequent subgraphs from graph databases [C] // Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, USA, 2004: 581-586
- [52] COOK J E, WOLF A L. Automating process discovery through event-data analysis [C] // 1995 17th International Conference on Software Engineering, Seattle, USA, 1995: 73-73
- [53] AGRAWAL R, GUNOPULOS D, LEYMAN F. Mining process models from workflow logs [C] // International Conference on Extending Database Technology: Advances in Database Technology, Valencia, Spain, 1998: 469-483
- [54] AALST W M P, DONGEN B F V, HERBST J, et al. Workflow mining: a survey of issues and approaches [J]. *Data & Knowledge Engineering*, 2003, 47(2): 237-267
- [55] AALST W M P, WEIJTERS T, MARUSTER L. Workflow mining: discovering process models from event logs [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2004, 16(9): 1128-1142
- [56] QIAN C, WEN L, KUMAR A, et al. An approach for process model extraction by multi-grained text classification [C] // International Conference on Advanced Information Systems Engineering, Grenoble, France, 2020: 268-282
- [57] DONGEN B F V, SMEDT J D, CICCIO C D, et al. Conformance checking of mixed-paradigm process models [J]. *Information Systems*, 2020: Doi: 10.1016/J.IS.2020.101685
- [58] TAX N, SIDOROVA N, HAAKMA R, et al. Mining local process models [J]. *Journal of Innovation in Digital Ecosystems*, 2016, 3(2): 183-196
- [59] DIAMANTINI C. Behavioral process mining for unstructured processes [J]. *Journal of Intelligent Information Systems*, 2016, 47(1): 5-32
- [60] WEN J F, ZHOU Z B, SHI Z S, et al. Crossing scientific workflow fragments discovery through activity abstraction in smart campus [J]. *IEEE Access*, 2018, 6: 40530-40546
- [61] HASSANI M, SICCHA S, RICHTER F, et al. Efficient process discovery from event streams using sequential pattern mining [C] // 2015 IEEE Symposium Series on Computational Intelligence, Cape Town, South Africa, 2015: 1366-1373
- [62] FAVER C, VÖLZER H, MÜLLER P. Diagnostic information for control-flow analysis of workflow graphs a. k. a. free-choice workflow nets [C] // International Conference on Tools and Algorithms for the Construction and Analysis of Systems, Eindhoven, The Netherlands, 2016: 463-479
- [63] GRECO G, GUZZO A, MANCO G, et al. Mining and reasoning on workflows [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2005, 17(4): 519-534
- [64] BOSE R P J C, AALST W M P. Abstractions in process mining: a taxonomy of patterns [C] // International Conference on Business Process Management, Ulm, Germany, 2009: 159-175
- [65] LEEMANS S S, FAHLAND D D, AALST W M P. Discovering block-structured process models from incomplete event logs [C] // International Conference on Applications and Theory of Petri Nets and Concurrency, Tunis, Tunisia, 2014: 91-110
- [66] SANI M F, AALST W M P, BOLT A, et al. Subgroup Discovery in Process Mining [M]. Cham: Springer International Publishing, 2017: 237-252

- [67] DALMAS B, TAX N, NORRE S. Heuristic Mining Approaches for High-Utility Local Process Models [M]. Transactions on Petri Nets and Other Models of Concurrency XIII, Berlin, Heidelberg: Springer Berlin Heidelberg, 2018: 27-51
- [68] TAX N, DALMAS B, SIDOROVA N, et al. Interest-driven discovery of local process models[J]. *Information Systems*, 2018, 77: 105-117
- [69] TAX N, SIDOROVA N, HAAKMAR R, et al. Mining local process models with constraints efficiently: applications to the analysis of smart home data[C] // 2018 14th International Conference on Intelligent Environments, Rome, Italy, 2018: 56-63
- [70] BUIJS J C M. Flexible evolutionary algorithms for mining structured process models [D]. Eindhoven: Eindhoven university of Technology, 2014: 36-38
- [71] LENO V, DUMAS M, MAGGI F M, et al. Automated discovery of declarative process models with correlated data conditions [J]. *Information Systems*, 2020, 89: 101482
- [72] PESIC M, SCHONENBERG H, AALST W M P. DECLARE: full support for loosely-structured processes[C] // Proceedings of the 11th IEEE International Enterprise Distributed Object Computing Conference, Washington, USA, 2007: 287-298
- [73] BERNARDI M L, CIMITILE M, FRANCESCOMARINO C D, et al. Using discriminative rule mining to discover declarative process models with non-atomic activities[C] // International Symposium on Rules and Rule Markup Languages for the Semantic Web, Prague, Czech Republic, 2014: 281-295
- [74] MANNHARDT F F, LEONI M M D, REIJERS H H, et al. Guided process discovery: a pattern-based approach[J]. *Information Systems*, 2018, 76: 1-18
- [75] GARIJO D. Mining Abstractions in Scientific Workflows [D]. Madrid: University of Southern California, 2015: 1-3
- [76] WANG J, JIN T, WONG R K, et al. Querying business process model repositories[J]. *World Wide Web*, 2014, 17(3): 427-454
- [77] GRECO G, GUZZO A, MANCO G, et al. Mining constrained graphs: the case of workflow systems[C] // Constraint-based Mining and Inductive Databases, Hintertzen, Germany, 2006: 155-171
- [78] JUNG J, BAE J, LIU L. Hierarchical business process clustering[C] // 2008 IEEE International Conference on Services Computing, Honolulu, USA, 2008: 613-616
- [79] DIAMANTINI C, GENGA L, POTENA D. Behavioral process mining for unstructured processes[J]. *Journal of Intelligent Information Systems*, 2016, 47(1): 5-32
- [80] GAMA M J. Conformance Checking and Diagnosis in Process Mining[M]. Cham: Springer International Publishing, 2016, 270
- [81] ROZINAT A, AALST W M P. Conformance checking of processes based on monitoring real behavior[J]. *Information Systems*, 2008, 33(1): 64-95
- [82] MUNOZ G J, CARMONA J, AALST W M P. Conformance checking in the large: partitioning and topology[C] // Proceedings of the 11th International conference on Business Process Management, Beijing, China, 2013: 130-145
- [83] POLYVYANYI A, ALKHAMMASH H, CICCIO C D, et al. Entropia: a family of entropy-based conformance checking measures for process mining[EB/OL]. <http://arxiv.org/pdf/2008.09558.pdf>; arXiv, (2020-09-30), [2021-03-01]
- [84] GAMA M J, CARMONA J, AALST W M P. Single-entry single-exit decomposed conformance checking[J]. *Information Systems*, 2014, 46: 102-122
- [85] CHEN C, YAN X F, ZHU F D, et al. gApprox: mining-frequent approximate patterns from a massive network[C] // The 7th IEEE International Conference on Data Mining, Omaha, USA, 2007: 445-450
- [86] LI R R, WANG W. REAFUM: representative approximate frequent subgraph mining[C] // Proceedings of the 2015 SIAM International Conference on Data Mining, Salt Lake City, USA, 2015: 757-765
- [87] SANI M F, ZELST S J V, AALST W M P. Improving process discovery results by filtering outliers using conditional behavioural probabilities[C] // International Conference on Business Process Management, Sydney, Australia, 2018: 216-229
- [88] 闫春钢, 蒋昌俊, 史有群, 等. 医疗信息系统工作流的 Petri 网建模与化简分析[J]. 高技术通讯, 2006, 16(6): 551-555
- [89] LIU C, ZENG Q, CHENG L, et al. Measuring similarity

- for data-aware business processes[J]. *IEEE Transactions on Automation Science and Engineering*, 2021, pp(99): 1-13
- [90] GOBLE C A, BHAGAT J, ALEKSEJEVS S, et al. my-Experiment: a repository and social network for the sharing of bioinformatics workflows[J]. *Nucleic Acids Research*, 2010, 38: 677-682
- [91] WOLSTENCROFT K, HAINES R, FELLOWS D, et al. The Taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud[J]. *Nucleic Acids Research*, 2013, 41: 557-561
- [92] TANG H, LIU Y, WANG L. A new algorithm of mining high utility sequential pattern in streaming data[J]. *International Journal of Computational Intelligence Systems*, 2018, 12(1): 342-350
- [93] FOURNIER V P, CHUN W L J, TRUONG C T, et al. High-Utility Pattern Mining: Theory, Algorithms and Applications [M]. Cham: Springer International Publishing, 2019: 1-45
- [94] GAN W, LIN C W, FOURNIER V P, et al. A survey of utility-oriented pattern mining[J]. *IEEE Transactions on Knowledge and Data Engineering*. 2021, 33(4): 1306-1327
- [95] THEIS J, DARABI H. Behavioral Petri net mining and automated analysis for human-computer interaction recommendations in multi-application environments[J]. *Proceedings of the ACM on Human-Computer Interaction*, 2019, 3(EICS): 13:1-13:16
- [96] LIU C. Automatic discovery of behavioral models from software execution data[J]. *IEEE Transactions on Automation Science and Engineering*, 2018, 15(4): 1897-1908
- [97] LU R, ZHANG S. BigCarl: mining frequent subnets from a single large Petri net[EB/OL]. <http://arxiv.org/pdf/2103.11342.pdf>; arXiv, (2021-03-21), [2021-05-21]
- [98] WANG M, DING Z, ZHAO P, et al. A dynamic data slice approach to the vulnerability analysis of e-commerce systems[J]. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020, 50(10): 3598-3612

Research progress of frequent subnets mining on workflow nets

ZHANG Shuhan^{* **}, FEI Chaoqun^{* **}, HUANG Xikun^{** **}, LI Yangyang^{**}

(^{*} Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences,
Institute of Computing Technology, Beijing 100190)

(^{**} Academy of Mathematics and Systems Science Key Laboratory of MADIS,
Chinese Academy of Sciences, Beijing 100190)

(^{***} University of Chinese Academy of Sciences, Beijing 100049)

Abstract

This paper summaries two main directions of frequent subnets mining on workflow nets, including constructing workflow nets and subnets from 1-dimension log processes data, and mining sub-structures from 2-dimension workflow nets. The advantages and disadvantages of these approaches are discussed. Due to the complex networked data structures and complete semantics that are contained in workflow nets, the issues of applying frequent pattern mining algorithms on workflow net mining are thoroughly analyzed. Furthermore, some typical applications of mining frequent subnets on workflow nets are presented, including anomalous pattern detection, cross-organizational variability analysis, and so on. Finally, some challenges and potential future directions concerning frequent subnets mining on workflow net are discussed.

Key words: frequent pattern mining (FPM), workflow net, subnet mining, Petri net, process mining