DOI:10.13196/j.cims.2015.02.002

基于 Artifact 操作日志的业务流程挖掘

王　颖[1,2],黄　震[1],刘国华[3]
(1.燕山大学 信息科学与工程学院,河北　秦皇岛　066004;
2.燕山大学 河北省计算机虚拟技术与系统集成重点实验室,河北　秦皇岛　066004;
3.东华大学 计算机科学与技术学院,上海　201620)

摘　要:为发现业务流程中的有用知识,结合以数据为中心的业务流程管理与流程挖掘技术,通过记录业务流程中关键数据实体的操作、操作者和操作时间,建立 Artifact 操作日志。基于 Artifact 操作日志进行流程挖掘,发现流程中的 Artifact 生命周期模型,分析流程中 Artifact 实例之间的依赖关系和执行者之间的社交网络关系,给出违规流程的查找算法。研究结果表明,以数据为中心进行业务流程挖掘能够发现关键业务数据之间的交互关系,为分析流程的正确性提供有效的方法。

关键词:以数据为中心;Artifact 操作日志;流程挖掘;流程社交网络
中图分类号:TP301　　**文献标识码**:A

Mining business process based on Artifact operation logs

WANG Ying[1,2], *HUANG Zhen*[1], *LIU Guo-hua*[3]
(1. School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China;
2. The Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province, Qinhuangdao 066004, China;
3. School of Computer Science and Technology, Donghua University, Shanghai 201620, China)

Abstract: To discover the useful knowledge in business process, by combining with the data-centric business process management and process mining technology, Artifact operation logs were established according to record the operation, operator and operate time of key data entity. Based on Artifact operation logs, Artifact lifecycle model was found with process mining, and the searching algorithm of deregulation process was given by analyzing the dependency relationship among Artifact instances and the social network among process executors. The result showed that the interactive relationship between key business data was discovered with data-centric business process mining, which could provide the effective method for analyzing the correctness of business.

Key words: data-centric; Artifact operation logs; process mining; process social network

0　引言

目前,以数据为中心已经成为业务流程管理的新的发展趋势,现有的主流业务流程建模工具如统一建模语言(Unified Modeling Language, UML)、业务流程建模标记(Business Process Modeling Notation, BPMN)和工作流网(WorkFlow-net, WF-net)等,都在流程描述中加入了表示数据对象的方法。20 世纪 90 年代,IBM 提出 Artifact 的概念[1],并围绕业务流程中的关键数据实体 Artifact 的生命周期变化进行流程建模。IBM 对以数据为中心的业务流程管理作了广泛的研究[2-6],并将研究成果应

收稿日期:2014-12-01。Received 01 Dec. 2014.
基金项目:国家自然科学基金资助项目(61472339)。**Founadation item**: Project supported by the National Natural Science Foundation, China (No. 61472339).

用于其以数据为中心的服务互操作项目(Artifact-Centric Service Interoperation,ACSI)[7]。在以 Artifact 为中心的业务流程系统中,Artifact 实例、对 Artifact 实例的操作和操作者形成了紧密的联系,如果能将流程运行时三者之间的关联关系抽取出来,则可以帮助流程管理者回答以下问题:哪些 Artifact 之间存在关联操作? 哪些人会操作同一个 Artifact 实例? 操作者之间是否存在不恰当的合作关系? 通过对流程运行结果的分析,将以上问题的答案通过可视化的形式显示,一旦流程出现问题,就可以快速找到与问题相关的 Artifact 实例和操作者,并且通过操作者之间的社交网络关系来发现流程中的违规操作。

大部分信息系统如工作流管理(WorkFlow Management, WFM)、企业资源计划(Enterprise Resource Planning, ERP)、客户关系管理(Customer Relationship Management, CRM)等都会在运行期间收集某种形式的日志数据以记录事件的执行,称为事件日志(event log)[8]。基于事件日志的流程挖掘是从事件日志中提取各种流程信息,包括流程模型、活动之间的时序关系、流程中的组织关系、流程执行者之间的社交关系等。在传统的以控制流为中心的业务流程管理领域,很多文献基于事件日志从不同角度进行流程挖掘,例如文献[9]从工作流事件日志中提取流程模型并用 Petri 网表示;文献[10]从事件日志中发现一种基于时序逻辑的约束模型;文献[11]基于日志发现同时带有控制流和数据条件的流程模型;文献[12]基于流程事件日志发现执行者的社交网络关系;文献[13]提出一种基于工作流日志挖掘的角色提取方法。但是,传统的日志是按照流程事件(case)分组的,当一个流程事件与多个数据实体相关联或者一个数据实体与多个流程事件相关时,不利于分析流程中关键数据的操作。在以 Artifact 为中心的业务流程管理领域,文献[14]从数据库中针对每一个 Artifact 相关事件提取日志,从而对以 Artifact 为中心的业务流程模型进行一致性检查。文献[15]从关系数据库中发现 Artifact 生命周期和 Artifact 之间的交互。然而以上工作没有针对以 Artifact 为中心建立的业务流程系统给出 Artifact 日志的结构并基于 Artifact 日志进行业务流程挖掘。

以 Artifact 为中心的业务流程管理在流程运行期间根据流程记录 Artifact 的操作日志,日志内容包括操作的 Artifact 实例、操作、操作者和操作时间。本文首先基于 Artifact 操作日志挖掘业务流程中的 Artifact 生命周期模型,然后将 Artifact 实例和操作者结合起来考虑,分析流程执行者之间的社交网络关系,揭示一定时间段内 Artifact 之间以及与操作者之间的关系,最后基于执行者社交网络图给出一个违规流程查找算法,并通过实例验证算法的有效性。

1　Artifact 操作日志

Artifact 是业务流程中的关键数据实体,用以描述完成一项业务所需的完整信息,很多业务中的订单、申请单都可以看作 Artifact。Artifact 有类型(type)和实例(instant)之分,类型是 Artifact 的结构,即 Artifact 所包含的属性及属性的值域,实例是对属性的具体赋值。下面分别给出定义。

定义 1　Artifact 类型。Artifact 类型 AT 是一个二元组(U,τ)。其中:

(1)U 是属性的有限集,有一个特殊的属性 $I\in U$ 称为标识符属性(identifier attribute)。

(2)$\tau:U\to D$ 是一个完全映射,D 是域(domain)的集合,D 中至少包含一个标识符域。

定义 2　Artifact 实例。设 AT 是一个 Artifact 类型,AT 的实例 a 是一个二元组(i,μ)。其中:

(1)μ 是一个部分映射,为 U 中的属性 u 分配域 $\tau(u)$ 中的元素。

(2)$i=\mu(I)$,且 $i\neq$NULL,称为标识符。

在流程执行过程中,会同时操作很多 Artifact 实例,表 1～表 4 所示为一个业务流程执行过程中记录的部分 Artifact 操作日志。Artifact 实例用 ID 属性做唯一标识,这里为了能够清晰地显示,对日志

表 1　Artifact 实例 BX_1 操作日志

Artifact 实例 ID	操作	操作者	操作时间
BX_1	活动 A	Wang	2014-04-16 9:10
BX_1	活动 B	Wang	2014-04-16 9:15
BX_1	活动 C	Zhao	2014-04-17 14:15

表 2　Artifact 实例 BX_2 操作日志

Artifact 实例 ID	操作	操作者	操作时间
BX_2	活动 A	Chen	2014-04-16 10:10
BX_2	活动 B	Wang	2014-04-17 8:30
BX_2	活动 C	Li	2014-04-17 9:30

表 3 Artifact 实例 BX_3 操作日志

Artifact 实例 ID	操作	操作者	操作时间
BX_3	活动 A	Chen	2014-05-25 15:10
BX_3	活动 B	Wang	2014-05-25 16:30
BX_3	活动 C	Li	2014-05-26 15:00

表 4 Artifact 实例 XM_1 操作日志

Artifact 实例 ID	操作	操作者	操作时间
XM_1	活动 D	Wang	2014-03-01 11:00
XM_1	活动 E	Zhang	2014-04-05 10:20
XM_1	活动 F	Liu	2014-04-15 10:30
XM_1	活动 C	系统	2014-04-17 9:31
XM_1	活动 C	系统	2014-04-17 14:16
XM_1	活动 F	Liu	2014-05-15 11:00
XM_1	活动 C	系统	2014-05-26 15:01
XM_1	活动 G	Wang	2014-06-10 10:20

进行了预处理，将日志中操作同一标识的 Artifact 实例的日志记录单独放在一张表中，从而得到 4 张表，分别记录 Artifact 实例 BX_1，BX_2，BX_3 和 XM_1 的操作日志。每张表中有 Artifact 实例 ID 列、操作列、操作者列和操作时间列 4 列。其中：Artifact 实例 ID 在同一张表中是相同的，操作列是操作 Artifact 实例的活动名，表中的行按时间顺序排列。

在表 4 中操作者“系统”是由系统自动完成的对 Artifact 实例的操作。活动 C 在完成对 Artifact 实例 BX_1，BX_2 和 BX_3 操作的同时，系统会自动操作 Artifact 实例 XM_1。

定义 3 Artifact 操作日志。设 A_i 是标识符为 i 的 Artifact 实例组成的集合，O 是 A_i 上操作的集合，P 是操作执行者的集合。Artifact 的操作日志 $L_i=\langle i,C^*\rangle$ 表示标识符为 i 的 Artifact 实例的操作日志，其中：$C=O\times P$ 是所有可能的行为，即操作及其执行者的组合；C^* 是按时间顺序排列的行为序列。

例如，表 3 的日志记录为〈BX_3，(活动 A，Chen)(活动 B，Wang)，(活动 C，Li)〉，表明 Artifact 实例 BX_3 由 Chen 通过活动 A 创建，然后转交给 Wang 完成活动 B，最后交给 Li 完成活动 C。

2 发现业务 Artifact 生命周期

生命周期是业务 Artifact 的重要特征，描述一类 Artifact 从创建到操作完成并归档的过程，生命周期的定义给出了业务逻辑正确实现的标准。当收集了充足的 Artifact 操作日志，即包含一类 Artifact 所有可能的操作时，就可以从日志中发现该类型 Artifact 的生命周期模型。IBM 最早在文献[1]中基于操作规范(Operation Specification，OpS)建立了 Artifact 生命周期模型。在文献[16]中，Richard Hull 等提出对 Artifact 生命周期建模的新方法——阶段里程碑(Guard-Stage-Milestone，GSM)，本文的 Artifact 生命周期采用 GSM 模型。

在第 1 章给出的 Artifact 操作日志中，表 1～表 3 对 Artifact 实例的操作均由活动 A，B，C 构成，说明 BX_1，BX_2，BX_3 是同一 Artifact 类型(假定类型名为 BX)的 3 个不同实例。表 4 给出的 Artifact 实例属于另一个 Artifact 类型(假定类型名为 XM)。

GSM 模型中，生命周期包括阶段(stage，用圆角矩形表示)、哨兵(guard，用菱形表示)和里程碑(milestone，用圆形表示)。阶段是完成某一目标的一组业务活动，阶段可以是原子的，也可以是复合的(包含其他子阶段)；每个阶段中有一个或多个哨兵，对应阶段的开始和终止，其中带十字的菱形表示创建新的 Artifact 实例的活动；里程碑表示重要的业务目标或者条件。

从表 1～表 4 的操作日志中，可以发现 Artifact 类型 BX 和 XM 的生命周期，将生命周期用 GSM 建模后如图 1 所示。BX 的生命周期有 3 个阶段，分别完成活动 A，B，C 制定的目标，每一阶段的开始以上一阶段的结束为前提。XM 的生命周期有 4 个阶段，其中第 1，2，4 阶段分别完成活动 D，E，G 的目标，第 3 阶段包括两个子阶段，分别对应活动 F 和活动 C，活动 F 和活动 C 在这一阶段可能交替执行。活动 C 在 BX 和 XM 的生命周期中对应两个不同的阶段，但是这两个阶段存在关联关系，如图 1 中的虚线箭头所示，BX 生命周期的阶段 3 目标完成后触发 XM 中子阶段的开始。

3 流程社交分析

Artifact 生命周期反映了业务流程中活动与活动之间的关联关系。基于 Artifact 操作日志，还可

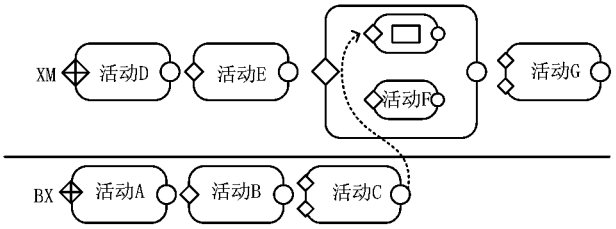

图1　GSM生命周期模型

以从另外的角度分析活动和执行者之间、执行者和执行者之间的关联关系。

3.1　发现流程角色

从 Artifact 操作日志中可以发现活动与执行者之间的关系:对于 Artifact 类型 BX 实例的操作,活动 A 由 Wang 或 Chen 完成,活动 B 由 Wang 完成,活动 C 由 Zhao 或 Li 完成;对于 Artifact 类型 XM 实例的操作,活动 D,E,F 和 G 分别由 Wang,Zhang,Liu 和 Wang 完成。因此,从角色分配上,Wang 和 Chen 是一个角色,Zhao 和 Li 是一个角色,Zhang 和 Liu 分别是一个角色。Wang 的角色不只一个,它参与多个流程,每个流程中都有不同的角色。

此外,还可以从执行者完成活动的时间角度进行分析。Artifact 实例的第一个活动执行者可以称为实例的创建者,最后一个活动执行者称为实例的归档者。例如 Wang 是 BX_1 和 XM_1 的创建者,Chen 是 BX_2 和 BX_3 的创建者;BX_1,BX_2 和 BX_3 在归档前都需要由 Wang 完成活动 B,并且 XM_1 的创建者和归档者都是 Wang,Wang 可能在 XM_1 中是负责人的角色。以上活动、角色和执行者之间可能的对应关系如图 2 所示。

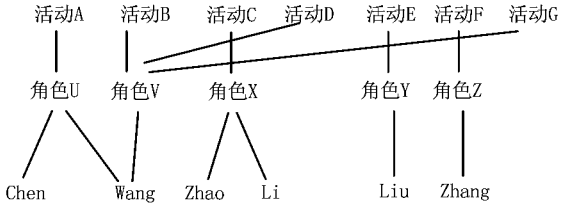

图2　活动、角色和执行者

3.2　发现社交网络

在对 Artifact 实例进行操作时,活动的执行者之间有先后承接的关系,即 Artifact 实例从一个执行者传递到下一个执行者。将具有承接关系的执行者相关联,可以构成一个执行者群体的社交网络图。

定义 4　执行者社交网络图 G 是一个三元组 $G(P,R,W)$。其中:P 为执行者的集合;$R=\{\langle p_1,p_2\rangle \mid p_1,p_2\in P\}$,表示 Artifact 实例从执行者 p_1 传递给 p_2;W 为权函数 $W(\langle p_1,p_2\rangle)\in N$,$N$ 是正整数集,表示 $p_1\sim p_2$ 的传递次数。

基于表 1～表 4 给出的 Artifact 操作日志,可以构造出图 3 所示的执行者社交网络图。图中忽略执行者是“系统”的情况,只考虑实际执行者之间的传递关系。

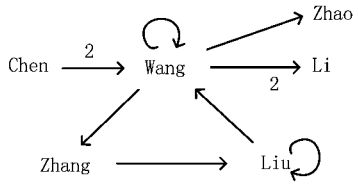

图3　执行者社交网络图

从生命周期分析中已经得出,Artifact 类型 BX 与 XM 是有依赖关系的,将日志中 Artifact 实例之间的依赖关系也用图的方式表达,称为 Artifact 实例依赖关系图。

定义 5　Artifact 实例依赖关系图 Q 是一个二元组 $Q(A,D)$,其中:A 为 Artifact 实例的集合;$D=\{\langle a_1,a_2\rangle \mid a_1,a_2\in A\}$,表示 Artifact 实例 a_1 和 a_2 的操作之间有依赖关系。

基于表 1～表 4 给出的 Artifact 操作日志,可以构造出图 4 所示的 Artifact 实例依赖关系图。

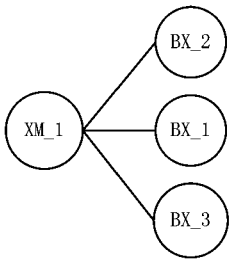

图4　Artifact实例依赖图

执行者社交网络图和 Artifact 实例依赖关系图均随时间变化,在流程执行的不同时间段,执行者社交网络图和 Artifact 实例依赖关系图的构成不同,假定时间区间 1,2,3 分别是 2014-03-01 8∶00 至 2014-04-16 12∶00、2014-04-15 8∶00 至 2014-04-17 12∶00、2014-04-17 14∶00 至 2014-06-10 12∶00,则各时间区间内的执行者社交网络图和 Artifact 实例依赖关系图如图 5 所示。

3.3　基于社交网络的分析

执行者社交网络图和 Artifact 实例依赖关系图提供了一个研究执行者之间、Artifact 实例之间、执行者和 Artifact 实例之间关系的基础,本文从以下几个角度进行分析。

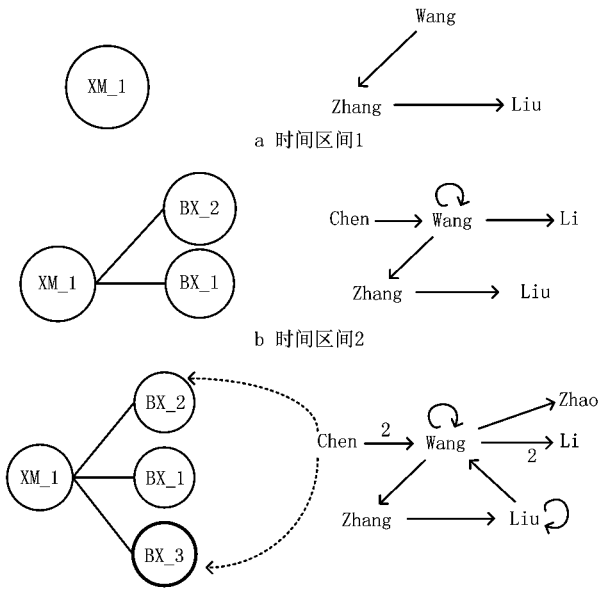

图5 不同时间区间的Artifact实例依赖图和执行者社交网络图

(1)Artifact 实例依赖关系图和执行者社交网络图的关联

Artifact 实例依赖关系图和执行者社交网络图之间是有关联关系的,Artifact 实例依赖关系图中的一个节点,可以对应执行者社交网络图中的多个节点,表示对一个 Artifact 实例进行操作的所有执行者。如图 5c 中的粗体部分对 Artifact 实例 BX_3 进行的操作分别有 Chen,Wang 和 Li。反之,执行者社交网络图中的一个节点可以对应 Artifact 实例依赖关系图中的多个节点,以显示一个执行者操作的所有 Artifact 实例。如图 5c 中的虚线箭头连接的部分所示,对 Chen 操作的 Artifact 实例有 BX_2 和 BX_3。

(2)查找违规流程

在业务流程执行过程中,有一些 Artifact 实例的传递过程是违规的。例如,如果 Chen 没有经过 Wang 直接将 Artifact 实例 BX_3 传递给 Li,则是一个违规流程。基于执行者社交网络图,通过子图查找,可以发现是否存在一些不正常的 Artifact 实例传递。如图 6 所示,在图 G 中查找是否存在子图 Q,如果存在,则说明流程执行中可能出现了违规实例。

算法 1 在执行者社交网络图中查找违规流程。

输入:执行者社交网络图 G,执行者传递关系图 Q。

输出:图 Q 是否存在于 G 中。

//图 Q 和 G 采用邻接表作为数据结构存储

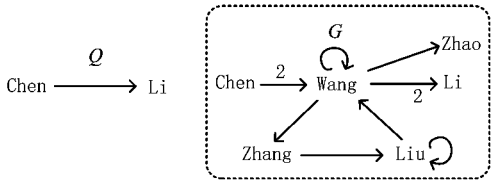

图6 违规流程查找

1. For each p∈P_Q do
2. 在 Q 和 G 的邻接表中分别找到以 p 为头节点的链表 L_Q 和 L_G;
3. 如果在 G 的邻接表中没有找到头节点 p
   则程序结束,图 Q 存在于 G 中,存在该违规流程;
   否则
4. For each q∈V_Q do /* 设 L_Q 中的节点集用 V_Q 表示 */
5. 在链表 L_G 中查找与节点 q 的值相等的节点;
6. 如果在 L_G 中没有找到与节点 q 的值相等的节点,则跳出内层和外层循环;
   /* 此时,说明图 Q 中至少有一条边不存在于图 G 中,则认为没有查找到该违规流程 */
7. 如果循环正常结束,则图 Q 存在于 G 中,G 中存在该违规流程,否则不存在

以图 6 所示的执行者社交网络图 G 和执行者传递关系图 Q 为例,图 Q 和 G 的邻接表如图 7 所示。根据算法 1 发现,图 Q 表示的从 Chen 到 Li 的传递关系在图 G 中不存在,即不存在该违规流程。

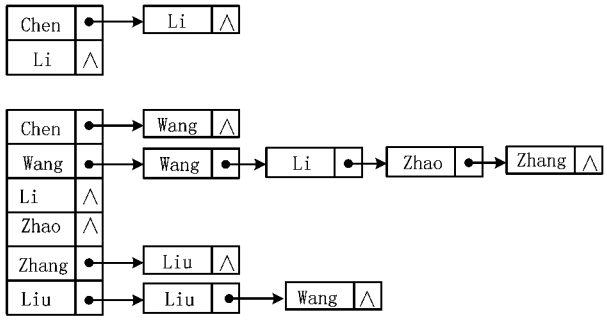

图7 图Q和G的邻接表存储结构

算法 1 是采用遍历的方式进行图的比较,算法分为两个阶段:①在图 G 中查找图 Q 中的节点;②在图 G 中查找与每一个节点关联的弧。假定图 Q 中有 m 个节点,图 G 中有 n 个节点,算法的基本操作是节点的比较,最坏情况下的算法复杂度为 $O(m\times(n+m\times n))$,即 $O(m^2\times n)$。

4 结束语

Artifact 是以数据为中心的业务流程管理中的核心概念,本文首先定义了 Artifact 操作日志的构成;然后基于日志进行流程挖掘,发现 Artifact 生命

周期模型、流程执行者社交网络图和 Artifact 实例依赖关系图;最后分析 Artifact 实例依赖关系图和执行者社交网络图的关联,并给出违规流程查找算法。本文将基于日志的流程挖掘和社交网络分析方法应用于以 Artifact 为中心的业务流程管理领域,下一步工作可以基于本文思想建立原型系统,作为模型一致性验证、业务流程监控和业务流程改进的工具。

参考文献:

[1] NIGAM A, CASWELL N S. Business artifacts: an approach to operational specification[J]. IBM Systems Journal, 2003, 42(3): 428-445.

[2] COHN D, HULL R. Business artifacts: a data-centric approach to modeling business operations and processes[J]. IEEE Data Eng. Bull., 2009, 32(3): 3-9.

[3] FRITZ C, HULL R, SU J. Automatic construction of simple artifact-based workflows[C]//Proceedings of International Confrence on Database Theory. New York, N. Y., USA: ACM, 2009: 225-238.

[4] CANGIALOSI P, DE GIACOMO G, DE MASELLIS R, et al. Conjunctive artifact-centric services[J]. Lecture Notes in Computer Science, 2010, 6470: 318-333.

[5] DAMAGGIO E, DEUTSCH A, VIANU V. Artifact systems with data dependencies and arithmetic[J]. ACM Transactions on Database Systems, 2012, 37(3): 2201-2236.

[6] VACULÍN R, HEATH T, HULL R. Data-centric Web services based on business artifacts[C]//Proceedings of the 19th International Conference on Web Services. Washington, D. C., USA: IEEE, 2012: 42-49.

[7] ARTALE A, LOMUSCIO A, HARIRI B B, et al. Artifact-centric service interoperation(ACSI)Web site[EB/OL]. (2010-06-01)[2014-11-02]. http://acsi-project.eu/.

[8] VAN DER AALST W M P, VAN DONGEN B F, HERBST J, et al. Workflow mining: a survey of issues and approaches[J]. Data and Knowledge Engineering, 2003, 47(2): 237-267.

[9] VAN DER AALST W M P, WEIJTERS A J M M, MARUSTER L. Workflow mining: discovering process models from event logs[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(9): 1128-1142,.

[10] MAGGI F M, BOSE J C, VAN DER AALST W M P. Efficient discovery of understandable declarative models from event logs[J]. Lecture Notes in Computer Science, 2012, 7328: 270-285.

[11] MAGGI F M, DUMAS M, GARCÍA-BAÑUELOS L, et al. Discovering data-aware declarative process models from event logs[J]. Lecture Notes in Computer Science, 2013, 8094: 81-96.

[12] VAN DER AALST W M P, SONG M. Mining social networks: uncovering interaction patterns in business processes[J]. Lecture Notes in Computer Science, 2004, 3080: 244-260.

[13] ZHAO Jing, ZHAO Weidong. Process roles identification based on workflow logs mining[J]. Computer Integrated Manufacturing Systems, 2006, 12(11): 1916-1920(in Chinese).[赵　静,赵卫东. 基于工作流日志挖掘的流程角色识别[J]. 计算机集成制造系统, 2006, 12(11): 1916-1920.]

[14] FAHLAND D, DE LEONI M, VAN DONGEN B F, et al. Behavioral conformance of artifact-centric process models[C]//Proceedings of the 14th International Conference on Business Information Systems. Berlin, Germany: Springer-Verlag, 2011: 15-17.

[15] LU X. Artifact-centric log extraction and process Discovery[D]. Eindhoven, the Netherlands: Eindhoven University of Technology, 2013.

[16] HULL R, DAMAGGIO E, DE MASELLIS R, et al. Business artifacts with guard-stage-milestone lifecycles: managing artifact interactions with conditions and events[C]//Proceedings of the 5th International Conference on Distributed Event-Based System DEBS. New York, N. Y., USA: ACM, 2011: 51-62.

作者简介:

王　颖(1980—),女,河北磁县人,副教授,博士,研究方向:业务流程管理、半结构化数据、Petri 网应用,E-mail:wangying@ysu.edu.cn;

黄　震(1976—),男,天津人,副教授,博士,研究方向:检测技术、图像处理;

刘国华(1966—),男,黑龙江依安人,教授,博士生导师,研究方向:数据库理论、数据库安全、Web 数据管理、业务流程管理。