第 33 卷 第 4 期 计 算 机 工 程 2007 年 2 月
Vol.33 No.4 Computer Engineering February 2007
·博士论文· 文章编号：1000—3428(2007)04—0020—03 文献标识码：A 中图分类号：TP391

工作流挖掘：一种新型工作流自动化建模方法

李 燕，冯玉强
(哈尔滨工业大学管理学院管理科学与工程系，哈尔滨 150001)

摘 要：为解决传统工作流建模方式主观性强、效率低以及成本高等问题，出现一种从工作流日志中自动化推导工作流模型的建模方法，又称工作流挖掘。这种新方法在国际上还处于研究初级阶段，但已引起广泛关注。该文对当前主要研究学者的工作及挖掘算法进行总结，介绍了两种不同类型挖掘算法的具体建模过程，并对其性能进行比较分析，用实例进行解释说明。

关键词：工作流建模；工作流挖掘；工作流日志；WF-net

Workflow Mining: Novel Automatic Method of Workflow Modeling

LI Yan, FENG Yuqiang
(Dept. of Management Science & Engineering, Management School, Harbin Institute of Technology, Harbin 150001)

【Abstract】 To solve the problems of strong subjectivity, poor efficiency and high cost in traditional workflow modeling, a novel modeling method automatically deriving model from workflow logs is appeared, which is called as workflow mining. At present this method is in primary stage in the world, but it has aroused extensive solicitude. This paper summarizes the scholars' main studies in workflow mining, introduces the modeling process of two different kinds of mining algorithms in detail, compares and analyzes their performances, and explains the modeling process with an actual example.

【Key words】 Workflow modeling; Workflow mining; Workflow logs; WF-net

工作流是一类能够完全或部分自动执行的经营过程[1]，它通过定义活动间的相互关系实现业务过程的自动化集成和管理。工作流建模通常由建模者根据企业工作流文档，或咨询流程参与者来定义模型，它简单、直观、易行，是当前建模的主要手段。然而该方式易受建模者个人能力及参与者主观经验的影响，即模型只表达出流程“应该”的样子，而非“实际”的样子，同时完全依赖建模者建模不仅效率低下且成本高。

对此，近年来一些专家提出一种新的工作流建模方式——基于过程日志的工作流自动化建模，又称为工作流挖掘（Workflow Mining）。这种建模方式采用数据挖掘思想，通过分析业务交互记录（即工作流日志），通过挖掘算法自动化推导企业工作流模型。目前大多数企业都建立了诸如 ERP，CRM，SCM 以及 WFMS 等系统。随着业务的不断执行，系统记录的工作流日志为企业的工作流挖掘提供了大量的数据基础。由这种建模方式建立的模型能反映出企业业务的实际执行过程，而且建模过程是自动化的，不仅可减少建模者及参与者对建模的影响，同时建模效率也可得到较大提高。

1 国内外研究现状

“工作流挖掘”思想最早出现在软件工程领域，是由美国新墨西哥州立大学计算机系教授Joanthan E.Cook于 1995 年提出的[2]。1998 年美国IBM Almaden研究中心的Rakesh Agrawal首次将其应用到企业工作流建模中[3]。

(1)Joanthan E.Cook 和 Alexander Wolf 在其文献[2]中设计了 3 种工作流挖掘方法，其中 KTail 和 Markovian 方法的实用性强于 RNet。但由于这 3 种方法皆用有限状态机(Finite-state Machine, FSM)作为工作流的形式化描述模型，因而受到 FSM 本身描述能力的限制，只能表示具有顺序结构的工作流模型。

(2)Rakesh Agrawal 等人在文献[3]中设计了两种工作流挖掘算法，并用“有向图”作为工作流的描述模型。这两种算法都可以推导出具有因果及循环关系的工作流模型，但却难以处理具有同步和选择关系的过程，即无法识别 And-join/split 和 OR-join/split 的过程逻辑结构。

(3)Gudio Schimm设计了一种可用于处理具有层次结构的工作流挖掘算法[4]，并开发出相应的挖掘工具（Process Minner)，但该算法难以推导出具有循环结构的工作流业务过程模型。

(4)Joachinm Herbst和Karagiannis将归纳法应用于工作流挖掘中[5,6]。但文献[5]的方法只能推导出具有顺序结构的工作流模型；文献[6]中的挖掘方法考虑到了过程同步问题，但却难于处理具有选择和循环关系的过程。另外，由于这两种挖掘方法推导的工作流模型只能用“随机活动图”表示，因此也具有一定的局限性。

(5)荷兰Eindhoven大学教授W.M.P van den Aalst是目前工作流建模领域的著名专家之一，他提出的工作流网WF-net建模理论与方法十分适用于企业业务过程建模。近年来，Aalst 致力于工作流挖掘技术与算法的研究与探索，并开发出相应的挖掘实验工具[7]。

Aalst 设计的挖掘算法主要分为两类：形式化挖掘算法和非形式化挖掘算法，且均用 WF-net 作为工作流挖掘结果的形式化表示模型。其中形式化算法要求工作流日志必须完整且

作者简介：李 燕(1979－)，女，博士生，主研方向：工作流管理，管理信息系统；冯玉强，教授、博导

收稿日期：2006-03-20 **E-mail**：liyuanyan7912@hit.edu.cn

不含噪音，且只能推导出结构合理的工作流模型，无法处理含有非自由选择结构的 WF-net；非形式化算法虽然对工作流日志无特殊要求，但却难于处理具有短循环结构的工作流模型。

2 相关概念及定义

在这一部份，我们将对工作流挖掘过程中用到的概念进行定义和说明。

定义 1(工作流日志) 设 T 为活动集合。$\sigma \in T^*$ 被称为工作流轨迹。$W \in P(T^*)$ 为工作流日志，其中 $P(T^*)$ 为 T^* 的集合。

定义 2(Petri 网) Petri 网 PN 是一个三元组(P,T,F):

· P 是有限个库所得集合；

· T 是有限个变迁的集合，$P \cap |\bullet t| = 1$；

· $F(P \times T) \cup (T \times P)$ 是弧的集合（流关系）。

定义 3(工作流网 WF-net) Petri 网 PN 被称为工作流网 N，当且仅当满足下面两个条件:

(1)PN 有两个特殊的库所: i 和 o。库所 i 是一个起始库所，即 $\bullet i = \Phi$；库所 o 是一个终止库所，即 $o\bullet = \Phi$。

(2)每一个节点 $x \in P \cup T$ 都位于从 i 到 o 的一条路径上。

定义 4(结构化工作流网 SWF-net) 工作流网 $N=(P,T,F)$ 是 SWF-net，当且仅当:

(1) $\forall p \in P$，$t \in T$，有 $(p,t) \in F$，则当 $|p\bullet| = 1$ 时，$|\bullet t| = 1$；

(2) $\forall p \in P$，$t \in T$，有 $(p,t) \in F$，则当 $|t\bullet| > 1$ 时，$|\bullet p| = 1$；

(3)不存在隐蔽性库所。

定义 5(活动间的基本关系) 设 W 为工作流日志，其中活动集为 T，有 W∈P(T*)，设 $a,b \in T$：

· 直接先后关系 $a >_w b$ 的充要条件：$\exists \sigma = t_1 t_2 \ldots t_{n-1}\ i \in \{1,\ldots,n\}, \sigma \in W$，有 $t_i = a$ 且 $t_{i=1} = b$

· 因果关系 $a \rightarrow_w b$ 的充要条件：$a >_w b$ 且 $a \not\succ_w b$ $_{ww}$；

· 选择关系 a#$_w$b 的充要条件:: $a \not\succ_w b$ 且 $b \not\succ_w a$；

· 并发关系 a// $_w$b 的充要条件：$a >_w b$ 且 $b >_w a$。

定义 6(∈, first,last) 设 T 为活动集合，$a \in T$ 且 $\sigma = a_1 a_2 \ldots a_n \in T^*$ 是长度为 n 的工作流轨迹。

· a ∈ σ 的充要条件：$a \in \{a_1 a_2 \ldots a_n\}$；

· $first(\sigma) = a_1$；

· $last(\sigma) = a_n$。

定义 7(完整的工作流日志) 设 $N=(P,T,F)$ 是合理的 WF-net:

· W 是 N 的工作流日志的充要条件：W ∈ P(T*)，且每组工作流轨迹 $\sigma \in W$ 都是 N 从初始状态[i]到结束状态[o]的点火序列，即为 $(N,[i])[\sigma > (N,[o])$

· W 是 N 的完整工作流日志的充要条件:

(1)对于 N 的任意工作流日志 W'，有 $>_{w'} \subseteq >_w$；

(2) $\forall t \in T, \exists \sigma \in W$，有 $t \in \sigma$。

3 工作流挖掘算法及比较分析

目前的工作流挖掘算法主要分为两种：形式化挖掘算法和启发式挖掘算法。形式化算法是通过形式化的方式从工作流日志中推导工作流模型；而启发式算法通过启发式规则推导日志中活动间的基本逻辑关系，建立工作流模型。

3.1 形式化挖掘算法——α 算法

α 算法通过形式化的方式识别工作流日志中活动间的基本关系，据此建立工作流模型，具体算法如下:

设 W 为工作流日志，T 是活动集合，$\alpha(W)$：

(1) $T_w = \{ t \in T \mid \exists_{\sigma \in w} t \in \sigma \}$

(2) $T_i = \{ t \in T \mid \exists_{\sigma \in w} t = first(\sigma) \}$

(3) $T_o = \{ t \in T \mid \exists_{\sigma \in w} t = last(\sigma) \}$

(4) $X_w = \{ (A,B) \mid A \in T_W \wedge B \subseteq T_W \wedge \forall_{a \in A, b \in B} a \rightarrow_W b \wedge \forall_{a1,a2} a1 \#_w a2 \wedge \forall_{b1,b2} b1 \#_w b2 \}$

(5) $Y_w = \{ (A,B) \in X_W \mid \forall_{(A',B') \in X_W} A \subseteq A' \wedge B \subseteq B' \Rightarrow (A,B) = (A',B') \}$

(6) $P_w = \{ p(A,B) \mid (A,B) \in Y_W \} \cup \{ i_w, o_w \}$

(7) $F_w = \{ (a,p(A,B)) \mid (A,B) \in Y_W \wedge a \in A \} \cup \{ (p(A,B),b) \mid (A,B) \in Y_w \wedge b \in B \} \cup \{ (i_w,t) \mid t \in T_I \} \cup \{ (t,o_w) \mid t \in T_o \}$

(8) α (W)=$\{ T_w$，P_w，$F_w \}$

α 算法的具体工作过程如下:

步骤(1)用于建立当前日志的活动集合 T_w；步骤(2)和步骤(3)建立输入变迁集 T_i 和输出变迁集 T_o；步骤(4)用于推导具有因果关系的变迁集合 X_w，其中集合 A 中的每个变迁与集合 B 中的每个变迁存在因果关系，但 A 和 B 中的变迁间不存在因果关系。通过对集合 A 与 B 的限制，可挖掘出 AND（OR）-Split/Join 工作流结构；在步骤(5)中通过建立 A 与 B 的最大集合重新定义 X_w；步骤(6)用于建立变迁间的库所；步骤(7)建立库所与变迁之间的流关系。通过步骤(1)~步骤(7)可从日志中推导出工作流模型，并用 WF-net 表示。

这种形式化的挖掘算法简单易懂，但在应用方面却存一定的局限性:

(1)该算法要求工作流日志必须完整，否则推导出的模型会丢失部分活动间的基本关系。显然形式化挖掘算法只适于结构简单且日志完整的工作流过程，对结构复杂、日志难于保证完整的工作流过程，算法实用性不强。

(2)由于日志中的噪音会影响活动间基本关系的推导，因此该算法的鲁棒性较低。

该算法难于识别含有 1 步和 2 步循环结构的工作流过程，只能建立 SWF-net，适应性不强。

3.2 启发式挖掘算法

启发式挖掘方法通过启发式规则从工作流日志中推导活动间的基本关系，分为 3 个步骤:

(1)创建依赖/频率表（dependency/frequenc able, D/F 表），如表 1 所示。

对于活动 a 和 b，可从日志中获得如下信息：1)活动 a，b 出现的总频率#a, #b；2)b 直接先于 a 的总频率#b<#a；3)b 直接继承 a 的总频率#a>#b；4)b 先于 a 的总频率#b<<<#a；5)a 先于 b 的总频率#a>>>#b；6)a，b 间的依赖关系强度值，记为 $\#a \rightarrow b$。

表 1 活动 b 对活动 a 的 D/F 表

#a	#b	#a>#b	#b<#a	#a>>>#b	#b<<<#a	$\#a \rightarrow \#b$

(2)根据 D/F 表推导活动间的基本关系。

定义启发式规则，并以 D/F 表为基础，推导活动间的基本关系（$a \rightarrow_w b$、$a \#_w b$ 或 $a //_w b$）。因果关系 a $\rightarrow_w$ b 的启发式规则：

IF ((#$a \rightarrow b \geq N$)) AND (#b<#$a \leq \theta$) AND (#a>#$b \geq \theta$) THEN $a \rightarrow_w b$

其中，N 是噪音影响因子，默认值 N=0.05，随着日志中噪音的增加而增大。θ 为域值且 $\theta = (1+ROUND(N \times \#L))/(\#T)$，其中#$L$ 代表日志中工作流轨迹的总行数，#T 代表总活动数。

(3)活动间的基本关系明确后，根据 α 算法建立工作流模型。

启发式算法改善了 α 算法的应用局限性。首先，由于在启发式算法中设置了参数 N 和 θ，因而可以有效处理含有噪音的日志，即算法具有鲁棒性。其次，由于该算法根据启发式规则推导活动间基本关系，故对日志的完整性没有严格要求。但由于挖掘前人们往往难以准确判断出日志中的噪音强度，因此会因为 N 和 θ 设置的不合理而影响挖掘的质量。

3.3 算法的比较分析

3.1 节和 3.2 节分别介绍了两类的工作流挖掘算法，它们有着各自的适用条件和应用领域，表 2 从不同的角度对其进行了比较和分析。

表 2 α 算法和启发式方法的比较分析

	使用条件	可挖掘的 WF-net 结构	对噪音的鲁棒性	工作流模型的结构	优点	缺点
α 算法	日志完整且无噪音	SWF-net（除 1 步和 2 步循环）	低	图表示	简单，易于理解，可操作性强	不适于结构复杂的工作流过程
启发式算法	无	SWF-net 和大部分 WF-net	强	图表示	适于挖掘各种复杂的工作流过程	N 和θ值不易确定

4 举例说明

设工作流日志（假设该日志完整且不含噪音）如表 3 所示，其中包含 5 个工作流实例 *case1-case5*，用 α 算法推导建立工作流模型。

表 3 工作流日志

案例 case	活动 activity	案例 case	活动 activity	案例 case	活动 activity	案例 case	活动 activity
Case1	A	Case1	C	Case5	E	Case4	B
Case2	A	Case2	C	Case4	C	Case5	F
Case3	A	Case4	A	Case1	D	Case4	D
Case3	B	Case2	B	Case3	C		
Case1	B	Case2	D	Case3	D		

(1)明确活动间的基本关系

日志中共含有 5 个工作流案例 σ_1={A,B,C,D}，σ_2={A,C,B,D}，σ_3={A,B,C,D}，σ_4={A,C,B,D}，σ_5={A,B,C,D,E}。根据定义 5 得：

·直接关系：A $>_w$ B，B $>_w$ C，C $>_w$ D，A $>_w$ C，B $>_w$ D，C $>_w$ B，E $>_w$ F；

·因果关系：A $\rightarrow_w$ B，C $\rightarrow_w$ D，A $\rightarrow_w$ C，B $\rightarrow_w$ D，E $\rightarrow_w$ F；

· 并发关系：B $//_w$ C，C $//_w$ B。

(2)用 α 算法进行工作流挖掘

1) T_w={A，B，C，D，E，F}；

2) T_i={A，E}；

3) T_o={D，F}；

4) X_w={（A，B）（C，D）（A，C）（B，D）（B，F）}；

5) $Y_w = X_w$；

6) P_w ={p(A,B)， p(C,D)， p(A,C)， p(B,D)， p(E,F)}$\cup$ { i_w , o_w }；

7) F_w ={（A，p(A,B)），(p(A,B)，B)，(C，p(C,D))，(p(C,D)，D)，(A，p(A,C))，(p(A,C)，C)，(B，p(B,D))，(p(B,D)，D)，(E，p(E,F))，(p(E,F)，F)，(i_w，A)(i_w，E)(D，o_w)，(F，o_w) }；

8) α （W）={T_w，P_w，F_w}。

(3)用 WF-net 描述工作流过程模型

工作流模型 WF-net 如图 1 所示。

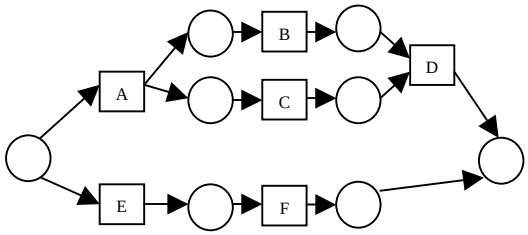

图 1 工作流模型 WF-net

5 结束语

本文总结工作流挖掘技术的研究现状，详细介绍了两种不同类型的挖掘算法——α 算法和启发式挖掘算法，并对其进行比较分析，指出它们各自的使用条件及优缺点，并用实例解释 α 算法的具体建模过程。

参考文献

1 范玉顺. 工作流管理技术基础[M]. 北京：清华大学出版社, 2001: 20-35.

2 Cook J E. Process Discoverying and Validation Through Event-data Analysis[R]. Boulder, University of Colorado, Technical Report: CU-CS-817-96, 1996-11.

3 Agrawal R, Gunopulos D, Leymann F. Mining Process Models from Workflow Logs[C]//Proceedings of the 6th International Conference on Extending Database Technology. 1998: 469-483.

4 Schimm G. Process Miner——A Tool for Mining Process Schemes from Event-based Data[C]// Proc. of the 8th European Conference on Artificial Intelligence. Springer-Verlag, 2002.

5 Herbst J, Karagiannis D. Integrating Machine Learning and Workflow Management to Support Acquisition and Adaptation of Workflow Models[J]. International Journal of Intelligent Systems in Accounting, Finance and Management, 2000, 9(5): 67–92.

6 Herbst J. Dealing with Concurrency in Workflow Induction[C]// Proc. of European Concurrent Engineering Conference on Society of Computer Simulation. 2000.

7 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, 9 (12): 369-378.