|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 流程挖掘算法与评估框架研究 | | | | | | | |
| (申请清华大学工学硕士学位论文) | | | | | | | |
|  | | | | | | | |
| 培 养 单 位 | | ： | 软件学院 | | |
| 学 科 | | ： | 软件工程 | |
| 研 究 生 | | ： |  | |
| 指 导 教 师 | | : |  | |
|  | | | | | | | |
|  | | 二○一五年五月 | | |  | | |

|  |
| --- |
| 流程  挖掘算法设计与评估框架 |
| 郭 |
| 秦 |
| 龙 |

|  |  |  |
| --- | --- | --- |
| **Study on the Design of Process Mining Algorithm and Evaluation Framework** | | |
| Thesis Submitted to  **Tsinghua University**  in partial fulfillment of the requirement  for the degree of  **Master of Science** | | |
| in  **Software Engineering** | | |
| by  **Guo Qinlong** | | |
| Thesis Supervisor | : |  |
|  | | |
| **May, 2015** | | |

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容。

本人保证遵守上述规定。

**（保密的论文在解密后遵守此规定）**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 作者签名： |  |  | 导师签名： |  |
| 日 期： |  |  | 日 期： |  |



摘 要

随着业务过程管理的蓬勃发展，许多大型企业和管理机构都使用业务流程模型来形式化内部流程。流程挖掘作为业务过程管理中一个关键技术，对事件日志数据分析，构造业务流程模型。流程挖掘技术可以应用于业务流程管理的许多场景，如业务过程模型的发现、一致性检查和改进。现存的流程挖掘算法无法同时挖掘非自由选择结构和不可见任务，导致无法胜任一些业务过程管理场景。与此同时，针对流程挖掘算法，至今已经有许多可以评估流程挖掘算法的评估框架，但是这些框架面临着训练数据训练选取复杂、描述模型的属性冗余的问题。围绕流程挖掘算法及其评估框架，本文提出了一个对挖掘不可见任务和非自由选择结构方面有独特优势的新流程挖掘算法，以及一个包含通用模型集合和约减的模型属性的评估框架。本文的主要工作如下：

1. 约减描述模型的属性：针对流程挖掘算法，提出对模型属性的两条约减规则。并且利用这两条约减规则将现有流程挖掘评估框架中的48个模型属性约减为6个。

(2)提出通用参考模型集合：根据6个描述流程挖掘模型属性，从人工模型中构造了一组通用的参考模型集合，这组通用参考模型可以解决现有流程评估框架中选择参考模型集合耗时与质量层次不齐的问题。

(3)设计了一个可以同时挖掘不可见任务与非自由选择结构的流程挖掘算法。该挖掘算法融合了两个现有的流程挖掘算法，并且解决了融合过程中带来的四个挑战性问题。

(4)针对约减的模型属性和通用参考模型集合设计实验，验证约减的模型属性和通用的参考模型集合对评估框架性能上的提升。并利用流程挖掘评估框架对本文提出的流程挖掘算法与现今主流流程挖掘算法评估，对比算法的效果。

关键词：业务过程管理；事件日志；工作流网；评估框架；流程挖掘；

**Abstract**

With the flourishing development of Business Process Management, lots of large companies and organizations are using process models to formalize their internal process. Process mining, as one of the key technology in Business Process Management, is a way to analyze event logs, and construct process models. Process Mining has been applied in to many scenarios in the Business Process Management: such as the discovery, conformance checking and enhancement of process models. Nowadays, existing process mining algorithms cannot mine process models with invisible tasks involved in the non-free-choice constructs, which made them incapable in supporting Business Process Management in some cases. Besides, there are some process mining evaluation framework to evaluate the process mining algorithms. However, it is time-consuming for these process mining algorithm evaluation framework to select the reference model set, and the quality of selected model set is not guaranteed. Besides, the features that characterizing process models in these framework has no relationship with the process mining aspect. This paper proposes a new process mining algorithm which is good at finding the invisible tasks involved in the non-free-choice constructs. Additionally, this paper proposes a universal significant reference model set and a reduced set of model features specializing on process mining. This paper contains work as follow:

1. Reducing process model feature: With focus on process mining, this paper proposes two criteria on reducing process model features. And by applying these two criteria, number of process model features in existing process model framework are reduced from 48 to six.

(2)Proposing a universal significant reference model set: According to the six reduced process model features, I construct a universal significant reference model set from artificial models. This model set is able to resolve the time-consuming and no-quality guarantee problem of selecting reference model set in existing process mining framework

(3)Designing a new process mining algorithm: This algorithm is able to mine process models with invisible task involved in non-free-choice constructs. This algorithm has combined two existing process mining algorithms, and solved four challenging issues involved in the combining.

(4) Design and conduct experiment on the reduced process model features and universal significant reference model set, with the result demonstrates the proposed features and model set improve the performance of process mining evaluation framework. Additionally, we use the improved process mining evaluation framework to compare our algorithm with existing mainstream process mining algorithms.

**Key words:** Business Process Management1; event log; workflow net; evaluation framework; process mining

目 录

[第1章 绪论 11](#_Toc415228107)

[1.1. 选题的背景与意义 11](#_Toc415228108)

[1.2. 基本概念 11](#_Toc415228111)

[1.2.1. 事件日志 11](#_Toc415228112)

[1.2.2. 工作流网 11](#_Toc415228113)

[1.2.3. 流程挖掘 11](#_Toc415228114)

[1.3. 论文的主要贡献 11](#_Toc415228116)

[1.4. 论文章节安排 11](#_Toc415228117)

[第2章 相关工作 12](#_Toc415228118)

[2.1. 相关挖掘算法介绍 12](#_Toc415228121)

[2.1.1. ALPHA算法 12](#_Toc415228125)

[2.1.2. ALPHA++算法 12](#_Toc415228126)

[2.1.3. ALPHA#算法 12](#_Toc415228127)

[2.1.4. 其他挖掘算法 12](#_Toc415228128)

[2.2. 流程挖掘算法评估框架介绍 12](#_Toc415228129)

[2.2.1. 基于专用模型全集的流程挖掘算法评估框架 12](#_Toc415228131)

[2.2.2. 基于专用模型子集与推荐技术的流程挖掘算法评估框架 12](#_Toc415228132)

[2.3. 本章小结 12](#_Toc415228133)

[第3章 面向不可见任务与非自由选择结构的流程挖掘算法设计 13](#_Toc415228134)

[3.1. 流程挖掘算法概述 13](#_Toc415228139)

[3.1.1. 研究动机 13](#_Toc415228140)

[3.1.2. 事件间基本关系 13](#_Toc415228141)

[3.1.3. 算法基本流程 13](#_Toc415228142)

[3.2. 发现改进的虚假依赖 13](#_Toc415228144)

[3.3. 补充可达关系 13](#_Toc415228145)

[3.4. 发现非自由选择结构 13](#_Toc415228146)

[3.5. 调整不可见任务 13](#_Toc415228147)

[3.5.1. 合并不可见任务 13](#_Toc415228162)

[3.5.2. 分离不可见任务 13](#_Toc415228163)

[3.6. 算法的实现 13](#_Toc415228164)

[3.7. 实验设计与分析 14](#_Toc415228165)

[3.7.1. 人工数据对比试验 14](#_Toc415228176)

[3.7.2. 真实数据对比试验 14](#_Toc415228177)

[3.8. 本章小结 14](#_Toc415228178)

[第4章 流程挖掘算法评估框架设计 15](#_Toc415228179)

[4.1. 流程挖掘算法框架整体介绍 15](#_Toc415228181)

[3.2. 流程模型的特征选取 15](#_Toc415228186)

[3.2.1. 特征选择标准 15](#_Toc415228187)

[3.2.2. 选取的特征 15](#_Toc415228188)

[4.3. 一组通用的重要参考模型集合 15](#_Toc415228189)

[3.3.1. 模型集合概述 15](#_Toc415228191)

[3.3.2. 不可见任务 15](#_Toc415228192)

[3.3.3. 重名任务 15](#_Toc415228193)

[3.3.4. 非自由选择 15](#_Toc415228194)

[3.3.5. 任意循环 15](#_Toc415228195)

[3.3.6. 短循环 15](#_Toc415228196)

[3.3.7. 嵌套循环 15](#_Toc415228197)

[4.4. 实验评估 15](#_Toc415228198)

[4.4.1. 对流程模型特征选取的评估 15](#_Toc415228206)

[4.4.2. 对参考模型集合的评估 15](#_Toc415228207)

[4.5. 本章小结 15](#_Toc415228208)

[第5章 总结与展望 16](#_Toc415228209)

[5.1. 总结 16](#_Toc415228211)

[5.2. 展望 16](#_Toc415228212)

主要符号对照表

第1章 绪论

* 1. 选题的背景与意义

随着业务流程管理技术的飞速发展， 越来越多的企业和组织引入各种各样的业务流程感知信息系统来支持其的业务流程[1]，包括WfMS[2][3][4] (如Staffware[5],IBM MQSeries[6]、COSA[7]）CRM软件，ERP软件（如SAP[8],Oracle[9]）。也有许多高校研发了许多业务流程的相关系统如BeehiveZ[10][11],Prom[12]。这些系统都可以有效支持业务流程管理中的建模、分析、仿真和优化。

流程感知系统的内部一般是通过流程模型来驱动的，然而创建流程模型是一项十分复杂与耗时的工作，而创建的流程模型往往可能是不完全的，主观的，过于抽象的或者与实际流程存在偏差，同时随着企业业务不断变更要求，流程模型也需要进行相应的变更。与此同时，有不少企业和组织为了提升服务质量，根据企业和组织的流程事件日志文件进行挖掘，产生了很多业务流程模型，从而更好的分析用户行为特点。流程挖掘技术就是为了解决以上困难而提出的。

最初，流程挖掘技术的思想是由来自美国新墨西哥州立大学的Cook教授在1995年提出[14]，它的目标是从表现软件流程的事件日志中自动地发现流程模型，并随之做了大量的后继工作，作者称该技术为流程发现。作者将流程模型表达为有限状态自动机，这样的流程模型能够表达并行结构和处理噪声。美国IBM艾曼登研究中心的Agrawal院士于1998年将Cook教授提出的流程发现技术引入业务流程领域，并正式命名为流程挖掘[15]。

目前国际上比较认可的一种流程挖掘的定义为“流程挖掘是指那些从实际的流程执行集合中提取出结构化的流程的描述的方法[15]”。

流程挖掘宣言[16]中将流程挖掘定义为“流程挖掘技术具有从现代信息系统中普遍产生的流程事件日志中抽取有价值信息的能力，该技术为各种应用领域中的流程发现、流程监测和流程改进提供了新的手段。”

流程挖掘的功能分为三种：发现、符合性检查、增强[17]。图1-1中展示了三种流程挖掘工程的输入和输出。发现是指对事件日志进行流程挖掘，通过流程挖掘，获得流程模型。可以看到图中，发现的输入是事件日志，输出是流程模型。符合性检查，是指通过一组事件日志和流程模型，判断流程模型和事件日志的记录是否相符合，可以看到在图中，符合性检查的输入是事件日志和模型，输出的是诊断结果。第三种功能是增强，增强是对于流程模型和在实际软件中产生的日志，对流程模型进行改进，提高流程模型执行的效率。在图中可以看到，增强功能的输入是事件日志和流程模型，输出是一个新的流程模型。

尽管到目前为止，已经有许多流程挖掘算法提出，但是仍然有不少挑战尚未解决[18]，如业务流程模型中的**非自由选择结构**与**不可见任务**同时存在于一个业务模型中的情况。非自由选择结构和不可见任务的具体含义详见第三章。目前的流程挖掘算法可以分别挖掘出非自由选择结构或者不可见任务结构，但是不具备挖掘同时含有这两种结构流程模型的能力。

与此同时，伴随着许多流程挖掘算法的提出，如何评价流程挖掘算法的能力成为了一个新的研究热点。来自荷兰埃因霍温理工大学的Rozinat对流程挖掘算法评估框架做出了开创性的研究[19]。他提出来的评估框架中共有四个模块：（1）事件日志和参考模型库，（2）事件日志生成模块，（3）修改和验证工具，（4）评估和比较模块。对于不同的挖掘算法，令其挖掘参考模型生成的事件日志，根据挖掘所得的模型和原始的参考模型作对比，挖掘所得模型与参考模型越相似说明流程挖掘算法挖掘能力越好。

但是Rozinat的流程挖掘算法评估框架存在一个重要的问题，就是评估需要讲所有的流程挖掘算法对所有的参考模型进行挖掘操作，然而一些流程挖掘算法本身的计算量是很大的，计算也需要花费很多的时间。因此一次评估框架进行一次评估往往需要很长的时间。针对这个问题，清华大学的王建民老师设计了利用线性回归的推荐思想设计新的评价框架。在新的评价框架中，参考的流程模型集合中选择一部分训练模型按照Rozinat的评估框架进行评估，然后根据评估结果和描述模型的48维向量计算线性回归模型，对于剩余部分的模型利用回归模型可以估算出应有的评估值，这样因为只针对参考模型集合中的一部分模型进行了流程挖掘，所以流程挖掘，所以这个新的评估框架的评估的时间得到了大幅的缩减。

然而这个利用线性回归的评估框架也存在一定的缺陷：首先，评估的效果收到选取的训练模型的质量，如果训练模型集合的质量不好，不能够全面的表达模型集合中的流程，那么训练出的线性回归模型会出现偏差，进而导致接下来的推荐效果出现误差。其次，在评估框架中采用了48维的模型属性来描述流程模型，但是这些属性中大部分与流程挖掘部分无关。

围绕流程挖掘中算法与评估框架，针对上面分析出的问题，本文的工作重点是设计了一个可以同时挖掘不可见任务和非自由选择结构的流程挖掘算法；以及设计一组通用的训练参考模型集合与约减描述流程模型的规则，改进了现有的流程挖掘框架。

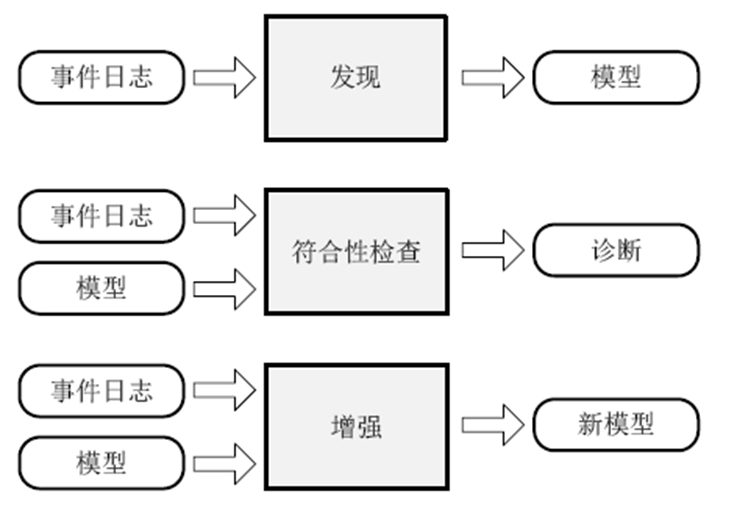


图 1.1 三种过程挖掘的输入输出

1. 2. 基本概念

本节主要介绍论文中使用的基本概念，其中1.2.1节介绍流程挖掘技术的输入，即事件日志，1.2.2介绍流程挖掘技术的输出，即流程模型与工作流网，1.2.3对流程挖掘技术进行介绍。

* + 1. 事件日志

事件，通常表示现实生活中的一个事件，是事件日志的基本单元。如同定义1中指出的，每个事件可以包含若干属性，比如说“时间戳”（即事件是什么时候发生的），“活动”（即事件所代表的活动内容），等等。

**定义1**（*事件, 属性*）令是时间空间，即表示所有可能的事件的集合，AN是一组事件属性的集合。对于事件集合的一个事件e（即），和AN中的一个属性n, 表示事件e在属性n下的值。

为了简化表达，我们认为规定了两个标准的事件属性：

* + 代表事件e的所指定的活动。
  + 代表事件e的对应的时间戳。

一个事件日志是由一组案例（case）组成的，每个案例（case）是由一组根据时间戳排好顺序的事件组成的，这组各不相同的事件表达形式为一个工作流轨迹（Trace）。对于每个案例，如事件一样，也可以由各种各样的属性。

**定义2**（*案例，轨迹，事件日志*）令是案例空间，即表示所有可能的案例的集合，AN是一组案例属性的集合。对于案例空间中的一个案例c，和AN中的一个属性n, 表示案例c对应的属性n的值。每个案例都有一个强制的属性trace,即表示这个案例对应的轨迹.通常，记 ，简化表达一个案例的轨迹。

一个轨迹是一个事件组成的有限序列，序列中的每个事件只出现一次。其中||表示轨迹的长度，即轨迹中包含的事件的个数。一个事件日志为一组案例组成的集合，即。

给定一组活动的集合T，一个定义在活动集合T上的事件日志W是指W中的任何一个事件的活动都在集合T里面，即有.

表1是一个事件日志的例子、这个事件日志中包含了四个案例。在本文中，我们仅考虑了事件的活动和时间相对顺序。例如案例1和案例4都表达了两个活动A和F，F在A之后执行。对于表1.1的案例，我们可以用一个简化的表达:{AF2,ADEG,ACDEGI}。其中AF的上标2表示AF组成的轨迹出现了两次。

表1.1一个事件日志的例子

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **案例编号(Case Id)** |  | **事件编号(Event Id)** | **属性** | |
|  | **时间戳(Timestamp)** | **活动(Activity)** |
| 1 |  | 35654422 | 30-12-2010:11.02 | A |
|  |  | 35654423 | 31-12-2010:13.32 | F |
| 2 |  | 35654481 | 01-01-2011:09.02 | A |
|  |  | 35654483 | 01-01-2011:13.24 | D |
|  |  | 35654484 | 01-01-2011:10.22 | E |
|  |  | 35654485 | 01-01-2011:21.22 | G |
| 3 |  | 35654579 | 03-02-2011:09.12 | A |
|  |  | 35654580 | 03-02-2011:09.12 | C |
|  |  | 35654581 | 03-02-2011:09.12 | D |
|  |  | 35654582 | 03-02-2011:09.12 | E |
|  |  | 35654583 | 03-02-2011:09.12 | G |
|  |  | 35654584 | 03-02-2011:09.12 | I |
| 4 |  | 35655442 | 30-12-2010:11.02 | A |
|  |  | 35655443 | 31-12-2010:13.32 | F |

* + 1. 工作流网

在本文中，我们使用工作流网来表示业务流程模型。

* + 1. 流程挖掘

  2. 论文的主要贡献
  3. 论文章节安排

第2章 相关工作

2. 1. 相关挖掘算法介绍
4. 1. 1. ALPHA算法
      2. ALPHA++算法
      3. ALPHA#算法
      4. 其他挖掘算法

**Genetic 算法**

**Heuristic 算法**

**Region 算法**

**ILP 算法**

* 1. 流程挖掘算法评估框架介绍
  2. 1. 基于专用模型全集的流程挖掘算法评估框架

**Rozinat**

**SAC 2012**

* + 1. 基于专用模型子集与推荐技术的流程挖掘算法评估框架

**ICWS 2012**

**TSC 2013**

* 1. 本章小结

第3章 面向不可见任务与非自由选择结构的流程挖掘算法设计

4. 1. 流程挖掘算法概述
      1. 研究动机

Motivating example.

* + 1. 事件间基本关系
    2. 算法基本流程

  2. 发现改进的虚假依赖
  3. 补充可达关系
  4. 发现非自由选择结构
  5. 调整不可见任务

4. 5. 1. 合并不可见任务
      2. 分离不可见任务
   6. 算法的实现

实现环境 prom/beehivez

实现界面

挖掘窗口、层叠日志和挖掘结果、体现特征

* 1. 实验设计与分析

3. 7. 1. 人工数据对比试验
      2. 真实数据对比试验
   8. 本章小结

第4章 流程挖掘算法评估框架设计

1. 1. 流程挖掘算法框架整体介绍

4. 2. 流程模型的特征选取
      1. 特征选择标准
      2. 选取的特征
   3. 一组通用的重要参考模型集合
   4. 1. 模型集合概述
      2. 不可见任务
      3. 重名任务
      4. 非自由选择
      5. 任意循环
      6. 短循环
      7. 嵌套循环
   5. 实验评估

7. 4. 1. 对流程模型特征选取的评估
      2. 对参考模型集合的评估
   5. 本章小结

第5章 总结与展望

1. 1. 总结
   2. 展望

参考文献

1. Dumas M, van der Aalst W M P, Ter Hofstede A H. Process-aware information systems: bridging people and software through process technology[M]. Wiley-Interscience, 2005..
2. van der Aalst W. M. P., van Hee K.M.. Workftow Managcment: Models, Methods and Systems. MIT press, Cambridge, MA, 2002
3. van der Aalst ,W. M. P., van Hee K.M.著.王建民 闻立杰 译.工作流管理—模型、方法和系统.清华大学出版社,北京,2004
4. Workflow Management Coalition. [EB/OL][2013-9-24] <http://www.wfmc.org>
5. TIBCO,Inc.Staffware Process Suite. [EB/OL][2013-9-23] <http://www.staffware.com>
6. IBM, Inc. IBM MQSeries [EB/OL][2013-9-23]. [http://www.mqseries.net/](http://www.mqseries.net/index.html)
7. COSA, Inc. COSA Business Process Management. [EB/OL][2013-9-25] <http://www.cosa-bpm.com>
8. SAP.SAP. [EB/OL][2013-9-23] <http://www.sap.com>
9. Oracle Business Process Suite. [EB/OL][2013-9-25]

<http://www.oracle.com/us/technologies/bpm/suite/overview/index.html>

1. 武年华,金涛,查海平等.BeehiveZ:一个开放的业务过程模型管理框架[J].计算机研究与发展,2010,47(z1):450c-454.
2. Jin T, Wang J, Wen L. Efficiently Querying Business Process Models with BeehiveZ[C]//BPM (Demos). 2011.
3. van Dongen B F, de Medeiros A K A, Verbeek H M W, et al. The ProM framework: A new era in process mining tool support[M]//Applications and Theory of Petri Nets 2005. Springer Berlin Heidelberg, 2005: 444-454.
4. J.E. Cook， A.L. Wolf. Automating Process Discovery Through Event-Data Analysis. Proceedings of the 17th international conference on Software engineering. New York， NY， USA，1995. 73–82
5. J.E. Cook， Z. Du， C. Liu， et al. Discovering models of behavior for concurrent workflows. Computers in Industry, 2004, 53(3):297–319
6. R. Agrawal， D. Gunopulos, F. Leymann. Mining Process Models from Workflow Logs. In: I. Ramos， G. Alonso， H.J. Schek， editors， Proceedings of Sixth International Conference on Extending Database Technology, 1998. 469–483.
7. W.M.P. van der Aalst， A.J.M.M.Weijters, and L.Maruster. Workflow Mining: Discovering Process Models from Event Logs. IEEE Transactions on Knowledge and Data Engineering， 16(9):1128–1142， 2004.
8. van der Aalst W M P. Process mining: discovery, conformance and enhancement of business processes[M]. Springer, 2011.
9. van der Aalst W M P, A.J.M.M. Weijters, Process mining: a research agenda, Computers in Industry, Volume 53, Issue 3, 2004, 231-244
10. Rozinat A, Medeiros A, Weijters A. The need for a process mining evaluation framework in research and practice. Computer Science, 2008:84-89.
11. van der Aalst, W. M. P., Weijters, T., Maruster, L. Workflow mining: Discovering process models from event logs. Knowledge and Data Engineering, IEEE Transactions on, 16(9), 1128-1142. (2004)