|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 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](#_Toc415147050)

[1.1. 选题的背景与意义 11](#_Toc415147051)

[1.2. 基本概念 11](#_Toc415147054)

[1.2.1. 事件日志 11](#_Toc415147055)

[1.2.2. 工作流网 11](#_Toc415147056)

[1.2.3. 流程挖掘 11](#_Toc415147057)

[1.3. 论文的主要贡献 11](#_Toc415147059)

[1.4. 论文章节安排 11](#_Toc415147060)

[第2章 相关工作 12](#_Toc415147061)

[2.1. 相关挖掘算法介绍 12](#_Toc415147064)

[2.1.1. ALPHA++算法 12](#_Toc415147068)

[2.1.2. ALPHA#算法 12](#_Toc415147069)

[2.1.3. 其他挖掘算法 12](#_Toc415147070)

[2.2. 流程挖掘算法评估框架介绍 12](#_Toc415147071)

[2.2.1. 基于挖掘结果比较的流程挖掘算法评估框架 12](#_Toc415147073)

[2.2.2. 基于推荐的流程挖掘算法评估框架 12](#_Toc415147074)

[2.3. 本章小结 12](#_Toc415147075)

[第3章 流程挖掘算法评估框架 13](#_Toc415147076)

[3.1. 流程挖掘算法框架整体介绍 13](#_Toc415147078)

[3.2. 流程模型的特征选取 13](#_Toc415147083)

[3.2.1. 特征选择标准 13](#_Toc415147084)

[3.2.2. 特征选取 13](#_Toc415147085)

[3.3. 一组通用的重要参考模型集合 13](#_Toc415147086)

[3.3.1. 模型集合概述 13](#_Toc415147088)

[3.3.2. 不可见任务 13](#_Toc415147089)

[3.3.3. 重名任务 13](#_Toc415147090)

[3.3.4. 非自由选择 13](#_Toc415147091)

[3.3.5. 任意循环 13](#_Toc415147092)

[3.3.6. 短循环 13](#_Toc415147093)

[3.3.7. 嵌套循环 13](#_Toc415147094)

[3.4. 实验评估 13](#_Toc415147095)

[3.4.1. 对流程模型特征选取的评估 13](#_Toc415147103)

[3.4.2. 对参考模型集合的评估 13](#_Toc415147104)

[3.5. 本章小结 13](#_Toc415147105)

[第4章 基于推理的面向不可见任务与非自由选择结构的流程挖掘算法研究算法概述 14](#_Toc415147106)

[4.1. 流程挖掘算法概述 14](#_Toc415147108)

[4.1.1. 事件间基本关系 14](#_Toc415147111)

[4.1.2. 算法基本流程 14](#_Toc415147112)

[4.2. 发现改进的虚假依赖 14](#_Toc415147113)

[4.3. 补充可达关系 14](#_Toc415147114)

[4.4. 发现非自由选择结构 14](#_Toc415147115)

[4.5. 调整不可见任务 14](#_Toc415147116)

[4.5.1. 合并不可见任务 14](#_Toc415147132)

[4.5.2. 分离不可见任务 14](#_Toc415147133)

[4.6. 算法的实现 14](#_Toc415147134)

[4.7. 实验设计与分析 14](#_Toc415147135)

[4.7.1. 人工数据对比试验 14](#_Toc415147147)

[4.7.2. 真实数据对比试验 14](#_Toc415147148)

[4.8. 本章小结 14](#_Toc415147149)

[第5章 总结与展望 15](#_Toc415147150)

[5.1. 总结 15](#_Toc415147152)

[5.2. 展望 15](#_Toc415147153)

主要符号对照表

第1章 绪论

* 1. 选题的背景与意义

1. 2. 基本概念
      1. 事件日志
      2. 工作流网
      3. 流程挖掘
   4. 论文的主要贡献
   5. 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

rozinate sac

* + 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 tsc

* 1. 本章小结

第3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4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

1. 1. 流程挖掘算法概述
2. 1. 1. 研究动机

motivating example.

* + 1. 事件间基本关系
    2. 算法基本流程
  1. 发现改进的虚假依赖
  2. 补充可达关系
  3. 发现非自由选择结构
  4. 调整不可见任务

5. 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

4. 7. 1. 人工数据对比试验
      2. 真实数据对比试验
   8. 本章小结
   9. 本章小结

第5章 总结与展望

1. 1. 总结
   2. 展望