Springer

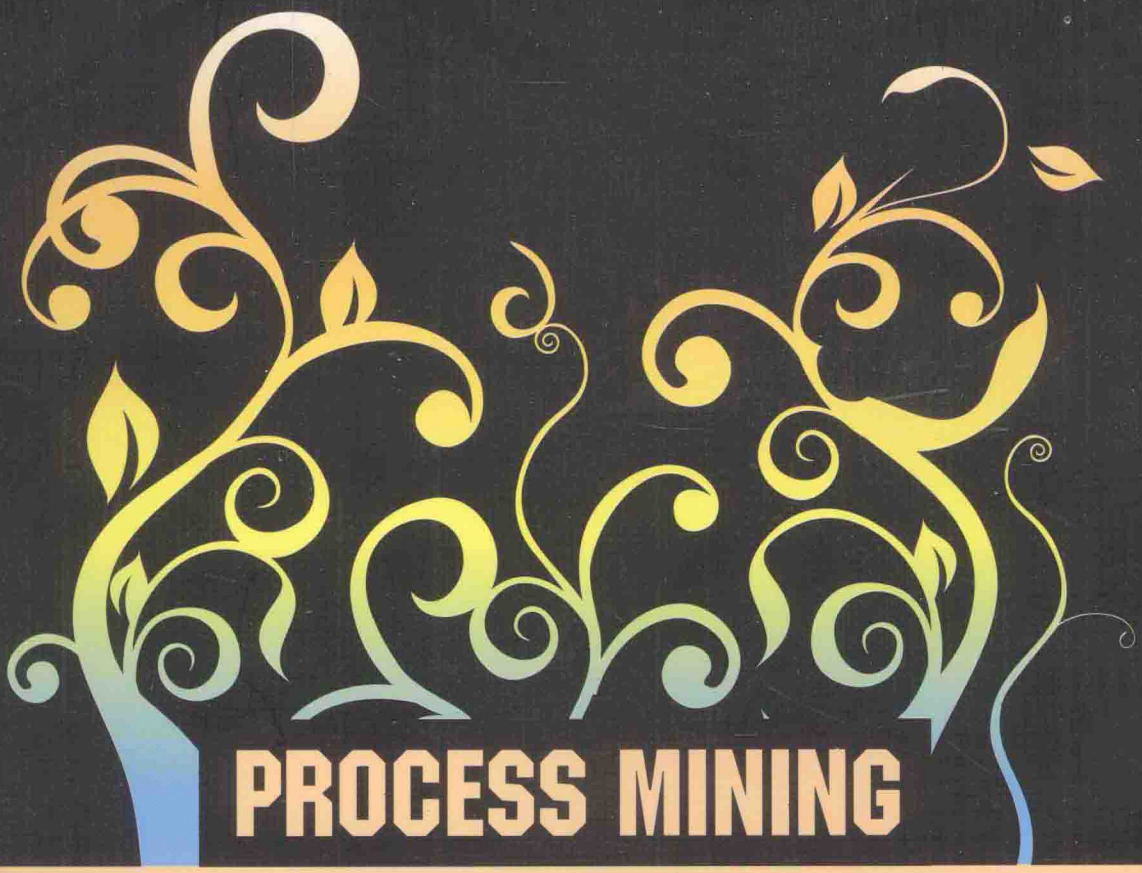
世界著名计算机教材精选

过程挖掘

业务过程的发现、合规和改进

Wil van der Aalst 著

王建民 闻立杰 等译



Discovery, Conformance and Enhancement of Business Processes

清华大学出版社

世界著名计算机教材精选

过程挖掘

业务过程的发现、合规和改进

Wil van der Aalst 著

王建民 闻立杰 等译

清 华 大 学 出 版 社

北 京

北京市版权局著作权合同登记号　图字：01-2012-7209 号

图书在版编目（CIP）数据

过程挖掘：业务过程的发现、合规和改进/（荷）阿尔斯特（Aalst, W.）著；王建民，闻立杰等译.
—北京：清华大学出版社，2014.
书名原文：Process Mining: Discovery, Conformance and Enhancement of Business Processes
世界著名计算机教材精选
ISBN 978-7-302-35085-9

Ⅰ. ①过…　Ⅱ. ①阿…　③闻…　Ⅲ. 数据采集–教材　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2014）第 009902 号

责任编辑：龙启铭
封面设计：何凤霞
责任校对：焦丽丽
责任印制：宋　林

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn, http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：三河市君旺印装厂
装 订 者：三河市新茂装订有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　印　　张：17.75　　字　　数：431 千字
版　　次：2014 年 6 月第 1 版　　印　　次：2014 年 6 月第 1 次印刷
印　　数：1～2000
定　　价：39.00 元

产品编号：044328-01

译 者 序

2011 年初荷兰埃因霍恩技术大学杰出教授、清华大学客座教授、荷兰皇家科学和人文学院院士、欧洲科学院院士、H-index 最高的欧洲计算机科学家 Wil 教授告诉我们，他将出版一本过程挖掘方面的书并希望再次合作[1]，从那时起我们就一直为这本书的中文译本努力工作并充满期待。

正如 IEEE 过程挖掘工作组在《过程挖掘宣言》中指出的：过程挖掘技术能够从现代信息系统普遍产生的事件日志中抽取过程知识，为相关领域应用中的过程发现、监测和改进提供了新的手段。

过程挖掘思想 1995 年起源于软件工程领域，1998 年被引入业务过程管理领域，是一个跨“数据挖掘”和“过程管理”领域的交叉学科。近十年来，伴随着事件数据获得越来越容易，过程挖掘技术快速发展，很多软件商已经将过程挖掘功能添加到其产品套件中。

过程挖掘主要场景包括：“发现”，根据事件日志生成模型，并不使用任何先验信息；“符合性检查”，将一个已知的过程模型与这个模型的事件日志进行对比；“改进”，使用相关事件日志来扩展或者改进现有过程模型。

过程挖掘并不限于过程发现，通过将事件数据和过程模型紧密联系，能用于检查合规性、探测偏差、预测延迟、支持决策制定和辅助过程再造，给原本静态的过程模型赋予了生机，并将现今的“大数据”置入过程上下文中。

本书理清了过程挖掘涉及的核心概念，重点阐述了事件日志获取方法、α 等控制流挖掘算法、合规性检查方法、数据/资源等属性挖掘方法、在线运作挖掘方法、过程挖掘项目框架以及过程挖掘典型领域案例，覆盖了从过程发现到运作支持整个过程挖掘技术谱系，最后对过程挖掘技术进行了展望。

清华大学软件学院信息系统与工程研究所的杨和东、朱笑尘、李婕、万明、窦蒙、沈晓明、林欣、冀付军、王子璇等同志参加了本书的翻译工作，在此，感谢他们的辛勤工作。

从 2011 年暑假我们得到书稿算起，已经两年有余。由于本书是过程挖掘领域的首部专著，涉及知识领域较多、组织视角独特，为了保证译文质量，译者认真研究、反复推敲，期望为读者奉献一个尽量准确的译文，这或许可以作为“两年有余”的一个理由。

最后衷心感谢 Wil 教授，他不仅让我们先睹为快，而且解答了翻译过程中的诸多疑问。

王建民　闻立杰

2014 年 3 月

于清华园

1　2004 年我们团队曾翻译 Wil 教授所著《工作流管理——模型、方法和系统》一书。

前　言

过程挖掘为改进各种应用领域中的过程提供了一种新的方式。这项新技术主要由两个因素驱动：一方面，越来越多的事件得以记录，提供了关于过程历史的详细信息。尽管事件数据无处不在，但大多数组织机构诊断问题时还是基于主观臆断而不是事实；另一方面，BPM（Business Process Management，业务过程管理）和 BI（Business Intelligence，商务智能）软件供应商一直都在大力推动过程挖掘技术。尽管 BPM 和 BI 技术得到广泛关注，它们与学者、顾问和软件供应商的期望尚有距离。

过程挖掘是一门提供全套工具来洞察事实并支持过程改进的新兴学科，这个新学科建立在过程模型驱动方法和数据挖掘的基础上。然而，过程挖掘并非现有方法的简单混合。例如，现有的数据挖掘技术主要以数据为中心，很难提供组织内端到端过程的全面理解。BI 工具则聚焦于简单的仪器盘和报表，而缺乏对商务过程进行清晰明确的洞察。BPM 套件则严重依赖于专家建模的理想化的未来过程，而无助于企业管理者理解现行的业务过程。

本书阐述了一系列过程挖掘技术，以帮助组织揭示它们的实际业务过程。过程挖掘并不限于过程发现，通过将事件数据和过程模型紧密联系，能用于检查合规性、探测偏差、预测延迟、支持决策制定和辅助过程再造。过程挖掘为原本静态的过程模型赋予了生机，并将现今的海量数据置入过程上下文中。因此，过程改进（如 Six Sigma、TQM、CPI 和 CPM）和合规性（如 SOX 和 BAM 等）相关的管理浪潮能够从过程挖掘中受益。

如书中所述，过程挖掘出现于十年前[8, 19]，但是它的起源却要追溯至半个世纪以前。例如，Anil Nerode 在 1958 年[101]提出从实例轨迹合成有限状态机的方法，Carl Adam Petri 则在 1962 年[103]介绍了第一门精确捕捉并发的建模语言，还有 Mark Gold 在 1967 年[67]率先系统地探索了不同的可学习的概念。当数据挖掘在 20 世纪 90 年代开始繁荣的时候，很少有人注意到过程挖掘。直到最近，事件日志变得无处不在，使得端到端的过程发现成为可能。自从关于过程挖掘的第一篇综述文章于 2003 年[8]发表后，过程挖掘取得了长足进步。过程挖掘技术日益成熟，出现了多种支持工具。尽管最初主要聚焦于过程发现，但现今过程挖掘谱系得到了明显的拓宽。例如，合规性检查、多维度过程挖掘和运作支持已成为过程挖掘工具——ProM——的有机组成部分。

这是关于过程挖掘的第一本书，因此，面向的读者群非常广泛。本书全面概述了过程挖掘发展现状，是为实践者、学生和学者提供的过程挖掘技术的入门材料。一方面，本书适合于那些刚接触该话题的新人；另一方面，本书对于重要概念也会给予严谨的阐述。本书力求自成体系，覆盖了从过程发现到运作支持的整个过程挖掘谱系。因此，对于 BPM 或 BI 领域的从业者，本书也可作为一本参考手册。

鉴于过程挖掘技术的适用性、（开源）过程挖掘软件的可用性，以及现代信息系统中丰富的事件数据，读者可以立即将过程挖掘技术付诸实践。我诚挚希望您乐于阅读本书，并着手使用那些已有的让人兴奋的过程挖掘技术。

Wil van der Aalst

致　谢

许多个人和机构都对本书中描述的技术和工具做出了贡献，在此衷心感谢他们的支持、付出和贡献。

本书源于 1999 年的研究项目“利用挖掘来进行过程设计：从执行日志中获得工作流知识”，这个项目是由本人和 Ton Weijters 发起的。那时，我还是美国科罗拉多大学波尔德分校的访问学者，BETA 研究院鼓励我与 TU/e（埃因霍恩技术大学）新研究组成员进行合作。经过和 Ton 讨论，觉得他机器学习方面的知识和我在工作流管理以及 Petri 网方面的知识会使我们的合作受益。显然，过程挖掘（当时我们称之为工作流挖掘）是一个能够将我们的专长进行结合的最合适的课题。于是就开始了这次成功的合作，感谢 Ton！

自此，许多博士生开始投身于这个课题，他们是 Laura Maruster、Ana Karla Alves de Medeiros、Boudewijn van Dongen、Minseok Song、Christian Günther、Anne Rozinat、Carmen Bratosin、R.P. Jagadeesh Chandra (JC) Bose、Ronny Mans、Maja Pesic、Joyce Nakatumba、Helen Schonenberg、Arya Adriansyah 和 Joos Buijs。我非常感谢他们的努力付出。

Ana Karla Alves de Medeiros 是第一位在我的指导下致力于这个课题的博士生。她做了非常杰出的工作，她关于遗传过程挖掘的论文获得了 ASML 2007 Promotion Prize 的奖项，并入选了 KNAW research school BETA 优秀论文。Boudewijn van Dongen 加入的时候，ProM 的开发才刚刚起步，作为一个硕士生，他开发了过程挖掘工具如 EMiT, 即 ProM 的前身，并成为一名优秀的博士生，在此期间他发明了多个过程挖掘技术。Eric Verbeek 在工作流验证方向完成了博士学位，随后他逐步参与到过程挖掘研究以及 ProM 开发中来。很多人低估了像 Eric 这样的学术型程序员的重要性。工具的开发和持续发展是科学研究的基石。Boudewijn 和 Eric 是 ProM 的推动力量，他们的贡献对于 TU/e 的过程挖掘研究是至关重要的。更重要的是他们一直助人为乐，谢谢你们！

Christian Günther 和 Anne Rozinat 是 2005 年加入团队的。他们对扩展过程挖掘的范围以及提升应用水平做出了重要贡献。Christian 在显著提高 ProM 性能的同时，使其更加美观。更重要的是，他的模糊挖掘插件促进了“意大利面过程”（Spaghetti）的处理。Anne 通过在 ProM 中加入合规性检查以及多维度挖掘，扩大了过程挖掘的应用范围。最重要的是他们建立了一家过程挖掘公司（Fluxicon）。Peter van den Brand 是另一位发展 ProM 的重要人物，他搭建了 ProM 6 的初始框架，并在其架构开发中发挥了重要作用。他基于 ProM 开发经验，建立了一家过程挖掘公司（Futura Process Intelligence）。与 Peter、Christian 和 Anne 这样的人共事非常愉快，他们对于将研究成果转化为商业化产品来说非常重要。我由衷希望 Fluxicon 和 Futura Process Intelligence 这两家公司能够持续成功发展（而不仅仅为了未来的跑车……）。

许多大学及其学者都对 ProM 的发展做出了贡献，并支持我们的过程挖掘研究。我们非常感谢 Technical University of Lisbon、Katholieke Universiteit Leuven、Universitat Politècnica de Catalunya、Universität Paderborn、University of Rostock、Humboldt-Universität

zu Berlin、University of Calabria、Queensland University of Technology、Tsinghua University、Universität Innsbruck、 Ulsan National Institute of Science and Technology、Università di Bologna, Zhejiang University、Vienna University of Technology、Universität Ulm、Open University、Jilin University、University of Padua 和 University of Nancy。我还要感谢 IEEE Task Force on Process Mining 的成员积极推动这个方向。我们也非常感谢以下机构对于 TU/e 过程挖掘研究的支持：NWO、STW、EU、IOP、LOIS、BETA、SIKS、Stichting EIT Informatica Onderwijs、Pallas Athena、IBM、LaQuSo、Philips Healthcare、ESI、Jacquard、Nuffic、BPM Usergroup 和 WWTF。特别致谢 Pallas Athena 对于过程挖掘课题的推动以及与他们多个项目的合作。有超过 100 个机构为我们提供事件日志，用以帮助提升过程挖掘技术。在此，我要特别提出 AMC Hospital、Philips Healthcare、ASML、Ricoh、Vestia、Catharina Hospital、Thales、Océ、Rijkswaterstaat、Heusden、Harderwijk、Deloitte 以及所有参加 SUPER、ACSI、PoSecCo 和 CoSeLoG 项目的机构。非常感谢他们让我们使用他们的数据并为我们提供反馈。

不可能列出所有对 ProM 有贡献的或帮助推动过程挖掘的每一个人，不过我还是要做一个尝试，除了之前提及过的人，我还要感谢 Piet Bakker、Huub de Beer、Tobias Blickle、Andrea Burattin、Riet van Buul、Toon Calders、Jorge Cardoso、Josep Carmona、Alina Chipaila、Francisco Curbera、Marlon Dumas、Schahram Dustdar、Paul Eertink、Dyon Egberts、Dirk Fahland、Diogo Ferreira、Walid Gaaloul、Stijn Goedertier、Adela Grando、Gianluigi Greco、Dolf Grünbauer、Antonella Guzzo、Kees van Hee、Joachim Herbst、Arthur ter Hofstede、John Hoogland、Ivo de Jong、Ivan Khodyrev、Thom Langerwerf、Massimiliano de Leoni、Jiafei Li、Ine van der Ligt、Zheng Liu、Niels Lohmann、Peter Hornix、Fabrizio Maggi、Jan Mendling、Frits Minderhoud、Arnold Moleman、Marco Montali、Michael zur Muehlen、Jorge Munoz-Gama、Mariska Netjes、Andriy Nikolov、Mykola Pechenizkiy、Carlos Pedrinaci、Viara Popova、Silvana Quaglini、Manfred Reichert、Hajo Reijers、Remmert Remmerts de Vries、Stefanie Rinderle-Ma、Marcello La Rosa、Michael Rosemann、Vladimir Rubin、Stefania Rusu、Eduardo Portela Santos、Natalia Sidorova、Alessandro Sperduti、Christian Stahl、Keith Swenson、Nikola Trcka、Kenny van Uden、Irene Vanderfeesten、George Varvaressos、Marc Verdonk、Sicco Verwer、Jan Vogelaar、Hans Vrins、Jianmin Wang、Teun Wagemakers、Barbara Weber、Lijie Wen、Jan Martijn van der Werf、Mathias Weske、Michael Westergaard、Moe Wynn、Bart Ydo 和 Marco Zapletal，感谢他们的支持。感谢所有阅读过这本书早期草稿的人（尤其是 Christian、Eric、Ton，感谢你们提出中肯的意见）。

感谢 Springer-Verlag 出版此书。感谢 Ralf Gerstner 鼓励我编写本书，并以非常出色的方式处理本书的出版事宜，谢谢你 Ralf!

本书 95%以上都是在漂亮的德国施莱登地区编写的。除了我的学术休假期间，平时还有许多其他任务需要处理。多亏我每周能来施莱登（这里没有因特网），从而可以在 3 个月内完成本书的编写。本书的逐章校对主要靠美妙的塞拉芬咖啡，其他写作时间主要靠这里美丽的风景。

按照惯例，最后的感谢要送给最珍贵的人。衷心感谢 Karin、Anne、Willem、Sjaak、Loes，在没有我的日子里，克服了很多困难。若没有她们持续支持，这本书也许会花费数年的时间。

目　录

第二部分 从事件日志到过程模型

第三部分 过程挖掘拓展

第四部分 过程挖掘的应用

第五部分 后 记

第1章 引 言

信息系统与它们所支持的运作流程越来越紧密地结合在一起，如今的信息系统记录了数量众多的事件。然而，企业很难从这些事件数据中提取有价值的信息。过程挖掘的目标就是从事件数据中提取过程相关的信息，比如，通过观察企业系统中的事件数据，自动地发现过程模型。为了说明过程挖掘的重要性，本章将讨论事件数据的急剧膨胀，及其对传统业务过程管理方法的挑战。然后，我们通过一个小例子来解释过程挖掘的基本概念。最后，阐述过程挖掘技术将为实现当代管理理念（如 SOX 和 Six Sigma）的目标，发挥重要作用。

1.1 数 据 爆 炸

信息系统和其他依赖于计算的系统，其拓展能力遵从摩尔定律。戈登·摩尔，英特尔公司的联合创始人，曾在 1965 年预测集成电路的元件数量每年都会翻番。在过去的 50 年中，这种增长确实是指数型的，只是步伐略有放缓。例如，集成电路上的晶体管数量每两年翻一番，磁盘容量、电脑的性价比、每一美元的像素数量等也都保持着类似的增长速度。除了这些令人难以置信的技术进步，人类和组织越来越依赖于计算机设备和网络上的信息源。2010 年 5 月 IDC 的“数字世界研究”显示出数据的惊人增长速度[79]。这项研究预测，数字信息（包括个人电脑、数码相机、服务器、传感器等）的存储量已经超过了 1 泽字节（Zettabyte），2010 年底这个“数字世界”将增长到 35 泽字节。IDC 的研究报告将 35 泽字节的数据描述成“一个高达地球到火星一半距离的 DVD 光盘堆”。这就是我们所说的数据爆炸。

从比特位到泽字节

“比特位”是信息的最小单位。比特位有两种可能的值：1（开）和 0（关）。一“字节”由 8 个比特位组成，可以表示 $2^8=256$ 个值。在描述大数据量时经常用 1000 的倍数：1 千字节（KB）等于 1000 字节，1 兆字节（MB）等于 1000KB，1 吉字节（GB）等于 1000MB，1 太字节（TB）等于 1000GB，1 拍字节（PB）等于 1 000TB，1 艾字节（EB）等于 1 000PB，1 泽字节（ZB）等于 1000EB。因此，1 泽字节是 10^{21}=1 000 000 000 000 000 000 000 字节。请注意，这里我们使用国际单位制（SI）单位前缀，又称 SI 前缀，而不是二进制前缀。如果我们使用二进制前缀，那么 1 千字节是 2^{10}=1024 字节，1 兆字节是 2^{20}=1 048 576 字节，1 泽字节是 $2^{70}\approx 1.18\times 10^{21}$ 字节。

绝大多数存储在“数字世界”中的数据是非结构化的，企业很难处理如此大量的数据，主要挑战之一就是如何从信息系统存储的数据中提取出有价值的信息。

“数字世界”和“物理世界”的融合日益深入，信息系统的重要性不仅体现在其中数据量的快速增长，同时还体现在这些系统对业务过程的有效支持。例如，“一个银行的状态”主要是由储存在这个银行的信息系统中的数据所决定的，货币已经被数字化了。当顾客在网络上预订机票时，他们会与多个组织进行交互（航空公司、旅行社、银行以及其他公司），而顾客往往并没有意识到这一点。如果预订成功，顾客将会得到一张电子客票。电子客票实际上是一串数字，它展示了数字世界和物理世界的融合。当一家大型制造商的 SAP 系统显示某件特定的产品已经没有库存时，即使在现实中这件货物有库存，它也无法被出售或是转运。诸如 RFID（Radio Frequency Identification，射频识别）、GPS（Global Positioning System，全球定位系统）与传感器网络这样的技术会进一步促进数字世界与物理世界的融合。RFID 标记使得每一个独立的物品都能被追踪。同时，我们注意到越来越多的设备正在被监控。例如，飞利浦医疗保健正在全球范围内监控它的医学设备（如 X 光机和 CT 扫描仪），这能够帮助飞利浦更好地了解客户需求、在实际环境中测试他们的系统、预见可能出现的问题、建立远程服务系统并从反复出现的问题中吸取经验。苹果公司“应用商店”的成功显示出，将地域感知和持续的互联网连接结合到一起的做法为无处不在的数字世界和物理世界的融合提供了新的途径。

与组织业务过程密切相关的数字世界的增长，使得记录和分析事件成为可能。事件包括：从 ATM 机中取一笔现金、一位医生设定 X 光机的剂量、一位市民申请驾照、提交一张纳税申报单，以及一名旅客收到一张电子客票。目前的挑战是如何更有效地利用事件数据，例如发现新知、找到瓶颈、预见问题、记录违反政策的行为、提出建议对策以及理顺流程等，所有这些都是过程挖掘所关心的问题！

1.2 建模的局限性

过程挖掘，即从事件日志中提取有价值的过程相关信息，是对现有业务过程管理（BPM）方法的补充。BPM 是一个学科，它结合了信息技术和管理科学的知识，并将其应用于运作业务过程[2,128]。近年来，由于 BPM 具有显著提高生产力和节约成本的潜力，得到了广泛重视。业务过程管理可以看成是 WFM（Workflow Management，工作流管理）的扩展。WFM 主要关注于业务流程自动化[10,80,85]，而 BPM 的范围更为广泛：从过程自动化、过程分析到过程管理和工作分配。一方面，BPM 的目标是改进那些可能未使用信息技术的业务过程，例如通过模拟仿真对业务过程进行建模和分析，管理层可能会获得通过提升服务水平来减少成本的思路；另一方面，BPM 常常通过软件来管理、控制和支持运作流程，这也是 WFM 的初衷。传统的 WFM 技术主要关注以“机械的”方式完成业务过程自动化，并不关注人为因素和对管理的深度支持。

过程感知的 PAISs（Process-Aware Information Systems，信息系统）包含传统的 WFM 系统，同时也包含能提供更大灵活性和支持特定任务的系统[58]，例如大型的 ERP（Enterprise Resource Planning，企业资源规划）系统（如 SAP、Oracle）、CRM（Customer Relationship Management，客户关系管理）系统、基于规则的系统、呼叫中心系统、高端中间件（如 WebSphere）等，均可以看作过程感知的信息系统，尽管它们没有必要使用通

用的工作流引擎来管理过程。事实上，这些信息系统都拥有明确的过程概念，也就是说，它们都能感知自己所支持的过程。数据库系统或电子邮件程序也会被用来执行某些业务过程中的某些步骤，但是这些软件工具无法“感知”用到它们的过程，即它们并不主动参与过程的管理和编排。一些作者使用术语 BPMS（BPM System，业务过程管理系统），或简单的使用 PMS（Process Management System，过程管理系统），来表达能够“感知”自己所支持的业务过程的系统。我们使用术语 PAIS 来强调它的范围远比传统的工作流技术要广。

BPM 和 PAIS 的共同点在于它们都依赖于过程模型。存在许多建模业务过程的表示语言（例如 Petri 网、BPMN、UML 和 EPC），第 2 章将对其中的一些语言进行讨论。这些表示语言的共同点在于都从过程所包含的活动（和可能的子过程）的角度来描述过程，这些活动的顺序由因果依赖关系决定。此外，过程模型也可以描述时间特性、指明数据的创建和使用（例如对决策建模），或者规定资源和过程的交互方式（例如角色、分配规则、优先级等）。

图 1.1 展示了一个用 Petri 网[52]表达的过程模型，这个模型描述了一家航空公司处理索赔申请的过程。顾客可能因为各种各样的理由要求索赔，例如航班晚点或者取消。如图 1.1 所示，该过程从注册一个申请开始，该活动用一个名为 register request 的变迁（用方框表示）来建模。变迁之间由库所连接，库所用圆圈来表示，用于建模过程的可能状态。在 Petri 网中，只有当一个变迁的所有输入库所中都含有托肯（token）时，这个变迁才是“使能的”（enabled），即该变迁所对应的活动是能够发生的。变迁 register request 只有一个名为 start 的输入库所，并且该库所最初就包含一个代表索赔申请的托肯，因此该变迁对应的活动是使能的。当“使能的”变迁“发生”（firing）时，它从自己的每个输入库所中消耗一个托肯，并在每个输出库所中产生一个托肯。也就是说，名为 register request 的变迁发生时会从 start 库所中消耗一个托肯，并产生两个新的托肯：一个在输出库所 c1 中，另一个在输出库所 c2 中。托肯用黑点表示，库所中托肯的分布（在本例中就是申请的状态）叫做标识（marking）。图 1.1 中的初始标识为库所 start 中包含一个托肯，其他库所没有托肯。在变迁 register request 发生后，标识中含有两个托肯，分别在输出库所 c1 和 c2 中。变迁 register request 发生后，有 3 个变迁处于使能状态。库所 c2 中的托肯使变迁 check ticket 使能，该变迁代表一项查看顾客是否有资格提出申请的行政检查，例如在检查机票的同时确认机票确实是由航空公司所发出的。与此同时，库所 c1 中的托肯使两个变迁 examine thoroughly 和 examine casually 同时使能。使能变迁 examine thoroughly 的发生，会移除库所 c1 中的托肯，从而使变迁 examine casually 不再使能。同样地，使能变迁 examine casually 的发生，会使 examine thoroughly 不再使能。换句话说，这两个活动之间是选择（有时也称为竞争）关系：当申请很复杂或可疑时，执行变迁 examine thoroughly，当申请比较简单时，只需要执行非正式的检查即可。执行 check ticket 并不会使其他变迁变得无法使能，即该变迁可以和变迁 examine thoroughly 或 examine casually 同时发生。变迁 decide 只有在它的所有输入库所都含有托肯时才能使能：票据需要被检查（库所 c4 中的托肯），对申请的简单复核或复杂复核也需要完成（库所 c3 中的托肯）。因此，在做决策之前整个过程同步了。变迁 decide 消耗掉两个托肯并在库所 c5 中产生一个托肯。有 3 个变迁共用 c5 中的托肯，代表着 3 种可

能做出的决策：支付所申请的赔偿（发生变迁 pay compensation），拒绝赔偿（发生变迁 reject request）或是需要更多的检查（发生变迁 reinitiate request）。对于最后一种情况，整个过程重新回到标识为 c1 和 c2 的状态：变迁 reinitiate request 消耗 c5 中的托肯并在其输出库所中产生托肯，这个标识刚好是变迁 register request 发生后的标识。原则上，可能发生多次迭代过程。整个过程在支付或拒绝赔款后结束。

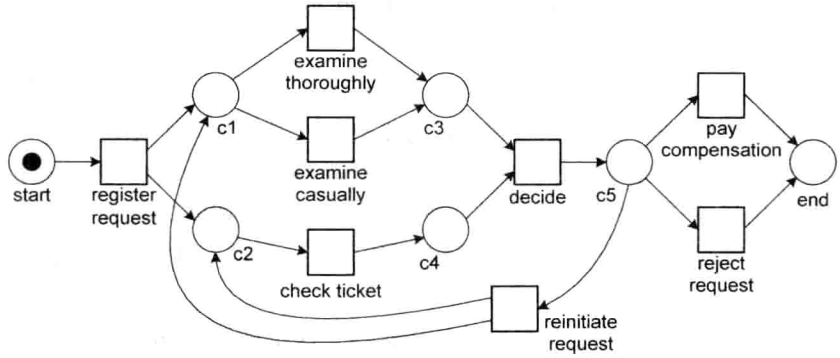


图 1.1 一个描述索赔申请处理过程的 Petri 网

图 1.1 将过程建模为一个 Petri 网，还有很多种其他的建模符号。图 1.2 利用所谓的 BPMN 图[102, 128]为同一个过程建模。BPMN（Business Process Modeling Notation，业务流程建模标注）不用库所，而使用显式的网关（gateway）去描述控制流逻辑。带有“×”符号的菱形代表 XOR 分裂和 XOR 合并，带有“+”符号的菱形代表 AND 分裂和 AND 合并。活动 register request 后面连接的菱形代表一个 XOR-合并网关，这个网关被用来表示在做出重审申请决定后的“跳回”操作。在 XOR-合并网关后是一个 AND-分裂网关，表示检查票据和简单/复杂审核是同时进行的。该 BPMN 图的其余部分也与之前的 Petri 网模型表示相同的行为。

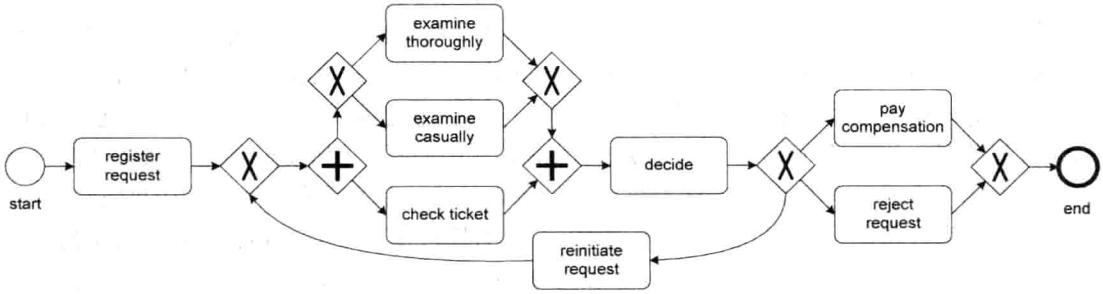


图 1.2 用 BPMN 表示的相同模型（索赔申请处理）

图 1.1 和图 1.2 只显示了控制流（control-flow），即所描述过程的活动发生顺序，这只是业务过程的一部分。因此，大多数建模语言都支持其他视角的描述符号，例如组织或资源视角（“决策需要由经理做出”）、数据视角（“除非申诉大于一百万欧元，否则决策总是在四次迭代内做出”）以及时间视角（“两周后问题将升级”）。虽然不同的过程建模语言之间存在很多差别，但这不是本书讨论的重点，我们在工作流模式[14, 131]相关研究中对此进行了系统比较。这里将重点阐述过程模型在 BPM 中所扮演的角色。

过程模型是用来做什么的?

- 洞察：在建模的过程中，引发建模者从不同的角度审视过程；
- 讨论：利益相关者可以使用模型来组织讨论；
- 文档化：过程被文档化，用于指导人们行动或明晰目标（参见 ISO 9000 质量管理）；
- 验证：分析过程以发现系统或程序中的错误（如潜在的死锁）；
- 性能分析：可以使用仿真等技术来发现对响应时间、服务水平等造成影响的因素；
- 模拟：模型能够使最终用户“播放”出不同的场景，从而为设计者提供反馈；
- 规约：模型可以在一个 PAIS 系统实现前描述它，因此可以作为开发者和最终用户/管理人员之间的“合同”；
- 配置：模型可以用来配置信息系统。

显然，过程模型在大型组织中扮演着重要的角色。当重新设计过程并引入新的信息系统时，过程模型会用于多种场景。通常会用到两类模型：（a）非形式化模型和（b）形式化模型（也叫做“可执行”模型）。非形式化模型用于讨论和归档，而形式化模型则用于分析和实施（即过程的实际执行）。前一种（a）对应于，示意高层次过程的“PowerPoint 图”；而后一种（b）对应于从运行代码中捕获的过程模型。非形式化模型通常是模糊的，而形式化模型往往关注点非常聚焦并且十分详细，便于利益相关者理解。这两种模型之间缺乏共性，这一点已经在 BPM 相关文献中被广泛讨论[2, 7, 58, 75, 78, 99, 128]。这里，我们提出另一种观点：与模型的种类——非形式化或形式化无关，一种反映模型与现实是否一致的观点。一个用于配置工作流管理系统的过程模型可能与现实保持良好的合规性，因为该模型会驱使人们按照其指定的方式去工作。不幸的是，大多数手工建立的模型都与现实严重脱节，并且只能提供手头过程的一种理想化视图。此外，允许进行严谨分析的形式化模型可能对实际过程帮助甚微。

模型如果不能很好地与现实联系起来，那么它的价值会十分有限。当与模型相关的人无法信任这些模型时，它们就变成了“纸老虎”。例如，对一个实际的过程而言，如果模型是真实过程的一个理想化版本，那么对这个模型进行仿真实验会毫无意义。这就好比是——基于一个理想模型——做出错误的重新设计的决定。按照一个忽视现实的过程模型去实现一个项目也是十分危险的。一个建立在理想化模型基础上的系统对于最终用户来说可能是毁灭性的和不可接受的，一个很好的例证就是大多数参考模型（reference model）的质量都十分有限。参考模型被用在大型企业系统中（例如 SAP），同时也被用于归档特定分支机构的过程，像 NVVB（Nederlandse Vereniging Voor Burgerzaken，即荷兰公民协会）的模型就描述了荷兰直辖市的核心过程。此处的想法在于让“最佳实践”在不同企业组织中共享。然而不幸的是，这些模型的质量与期望相去甚远。例如，SAP 参考模型对 SAP 实际支持的过程帮助非常少，事实上，超过 20%的 SAP 模型含有严重的瑕疵（死锁、活锁等）[95]。这些模型与现实脱节，并对最终用户价值甚微。

鉴于（a）对过程模型的兴趣（b）大量的事件数据以及（c）手工模型有限的质量，将事件数据与过程模型结合起来将是很有意义的。采用这种方法，实际的过程可以被发现，

现有的过程模型也能够被评估和改进，这正是过程挖掘致力于达到的目标。

1.3 过 程 挖 掘

为了定位过程挖掘，我们首先用图 1.3 描述 BPM 生命周期这一概念。生命周期描述了管理一个特定业务过程的不同阶段。在设计阶段，过程模型被设计出来。在配置/实现阶段，这个模型被转换成一个可运行系统。如果模型已经处于可执行的形式，并且一个 WFM 或 BPM 系统已经在运行，这个阶段可能会非常短暂。但是，如果这个模型是非形式化的并且需要被硬编码为传统的软件形式，这个阶段可能会耗费大量时间。在系统支持所设计的过程之后，实施/监控阶段开始。在此阶段，过程将在管理者的监控下运行，以便发现是否需要修改。其中一些修改将在图 1.3 的调整阶段进行处理。在调整阶段，过程不会被重新设计，也不会生成新的软件；只有预定义的控制会被用来适应或重新配置过程。诊断/需求阶段评估过程并监控不断出现的需求，以应对过程所在环境出现的变化（例如政策、法律、竞争的变化）。欠佳的表现（如不能达到服务水平）或环境造成的新变化会触发 BPM 生命周期从重设计阶段开始新一轮的迭代。

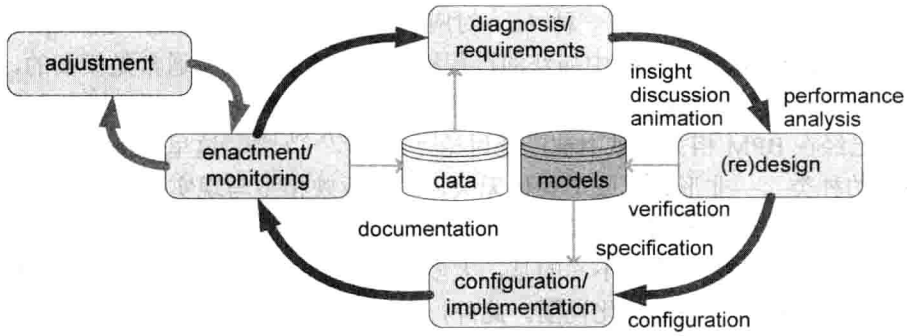


图 1.3 BPM 生命周期显示了过程模型的不同用途

如图 1.3 所示，过程模型在（重）设计和配置/实现阶段扮演主导角色，而数据在实施/监控阶段和诊断/需求阶段扮演主导角色。本图还列出了过程模型的不同用途（在 1.2 节中定义）。到目前为止，过程执行和实际过程设计二者产生的数据之间的联系很少。事实上，在大部分组织中，诊断/需求阶段都未被系统和连续的方式加以支持。只有严重问题或主要外部变化会激发生命周期的新一轮迭代，同时有关当前过程的实际信息并没有积极地被用于重定义的决策中。过程挖掘提供一种真正“闭合”BPM 生命周期的可能性。信息系统记录的数据可以被用来提供一个更好的关于实际过程的视图，也就是说，偏差可以被分析并且模型的质量能够得到提高。

过程挖掘是一门相对年轻的研究学科，它一方面位于机器学习和数据挖掘之间，另一方面又位于过程建模与分析中。过程挖掘的理念是通过从事件日志中提取出知识，从而去发现、监控和改进实际过程（即非假定的过程），而事件日志在如今的系统中是很容易获得的。

如图 1.4 所示，过程挖掘建立了两种连接，一是实际过程与其数据的连接；二是实际