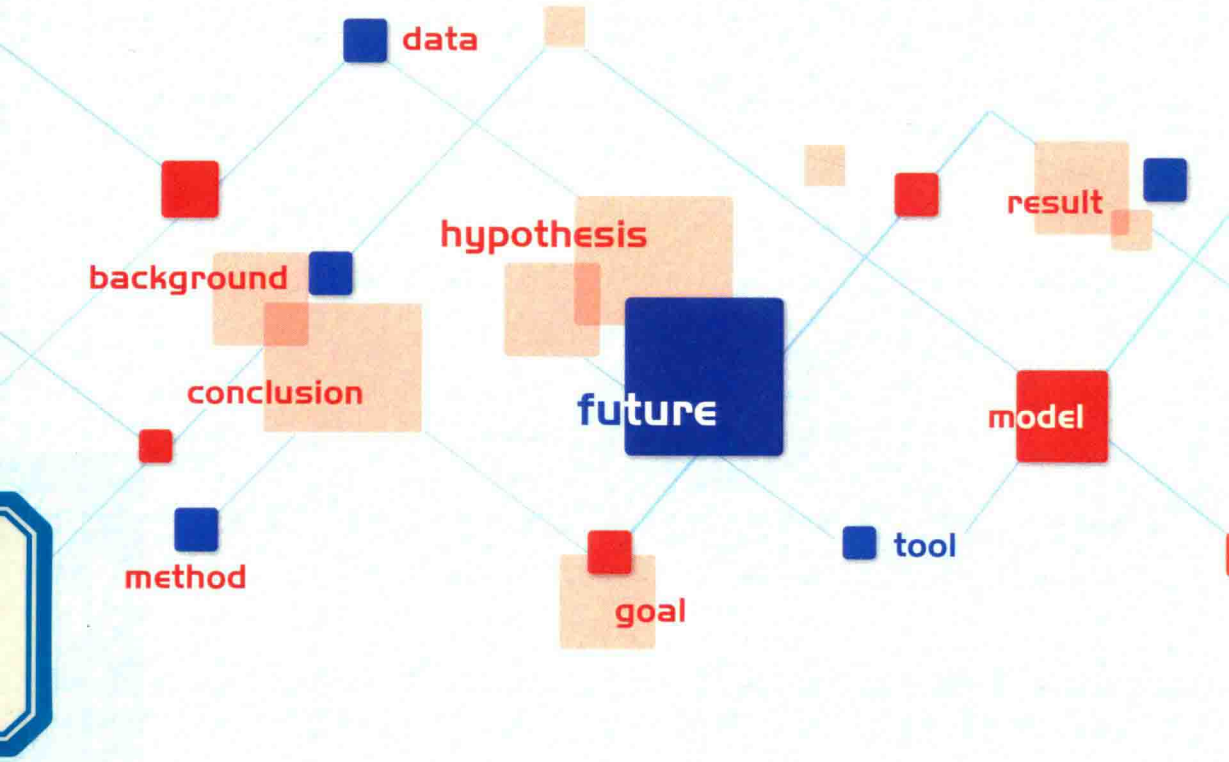
钱力 张晓林 著

科技论文的研究设计

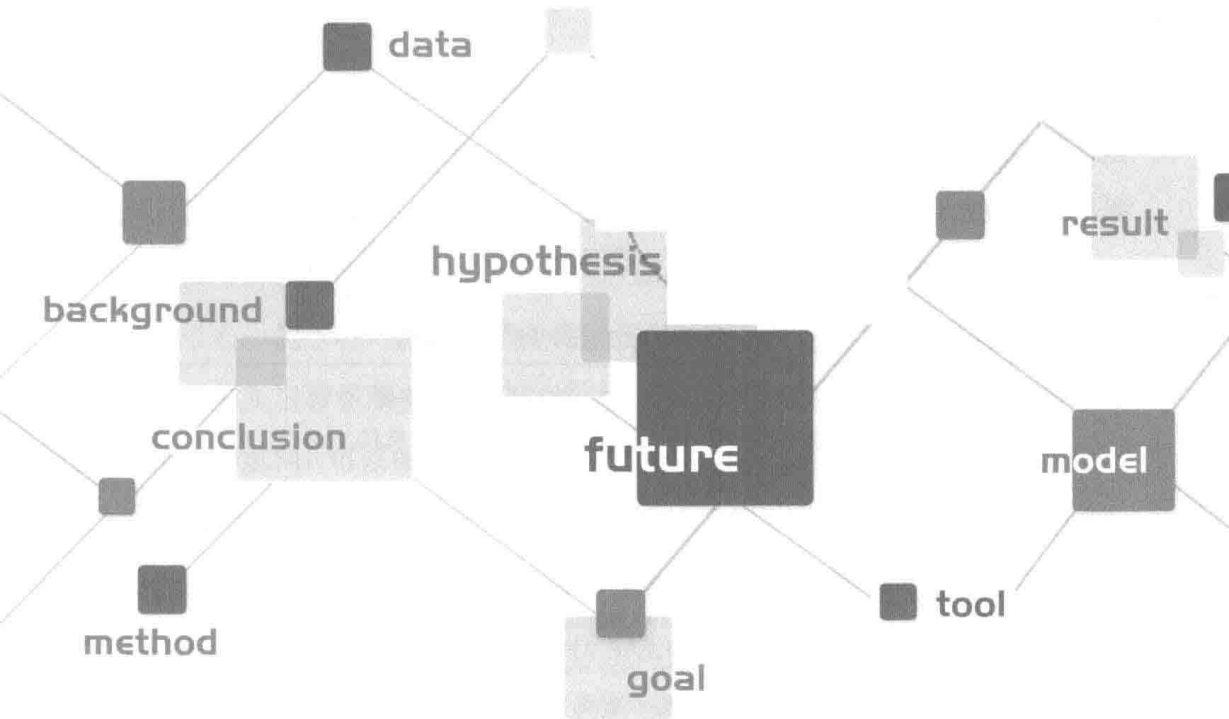
指纹识别方法研究



科学出版社

钱力 张晓林 著

科技论文的研究设计
指纹识别方法研究



科学出版社
北京

内 容 简 介

在大数据与人工智能快速发展的背景下，科技情报知识服务向智慧化、精准化的转型升级成为现阶段的迫切需求，而面向科技论文的细粒度知识抽取、语义关联与知识计算是这一转型升级过程中的关键核心。本书从科技学术论文的研究过程出发，提出了能够结构化描述科技论文核心知识的研究设计指纹框架，包括研究问题、研究目标、研究方法、研究工具及研究结论等指纹特征，并基于知识库与深度学习技术方法，实现了研究设计指纹特征的自动识别与示范应用。相关的研究方法与研究成果，为文献情报下一代智能化、个性化的精准知识服务提供基础的数据智能化计算支撑。

图书在版编目（CIP）数据

科技论文的研究设计指纹识别方法研究/钱力，张晓林著. —北京：科学出版社，2019.3

ISBN 978-7-03-058838-8

Ⅰ. ①科… Ⅱ. ①钱… ②张… Ⅲ. ①指纹鉴定-方法研究 Ⅳ. ①D918.91

中国版本图书馆 CIP 数据核字（2018）第 210334 号

责任编辑：徐 烁/责任校对：贾娜娜

责任印制：张 伟/封面设计：楠竹文化

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

北京建宏印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2019 年 3 月第 一 版 开本：720×1000 B5

2019 年 3 月第一次印刷 印张：10 1/2

字数：143 000

定价：58.00 元

（如有印装质量问题，我社负责调换）

前　言

目前，人工智能正引发链式反应般的科学突破，引领新一轮科技革命和产业变革，而作为支撑人工智能发展的科技论文数据，记载着科学真理验证过程、实验观测、研究结论、网络交流等科技情报知识线索，是人工智能用于科技创新发现的算法模型实现的数据根基和知识基础，而且以语义化知识数据为基础的知识服务及其“人—机—物”三元计算体系，已经成为 Google、Microsoft 等企业抢占未来大数据人工智能服务的重要部署。在此基础上，基于语义数据的科技知识的深入挖掘和重构，可以促进前沿识别、颠覆性技术识别和技术交叉前沿发现等科研创新，同时也使知识体系不断丰富化、细粒度化和语义化。因此，从多类型、多层级、多粒度上，计算挖掘出科技论文中的核心知识内容变得十分必要与迫切。

本书围绕上述发展趋势，在调研分析科技论文核心知识内容识别与抽取的相关技术基础上，提出、设计并实现了面向科技论文的研究设计指纹识别模型与方法体系。该体系以研究设计指纹为切入点，重点探讨了研究方法、研究工具、研究结论等指纹特征知识的自动识别机制，实现了针对科研设计的研究方案挖掘（Solution Mining）。

本书共分七章。第一章，主要提出并阐述研究设计指纹的内涵与类型，并总体介绍研究设计指纹识别的研究思路与方法。其中，研究设计指纹分为基础指纹、

技术指纹、结论指纹与未来指纹四个大类，包括背景指纹、方法指纹、工具指纹等九种类型，为有效地实现对科技论文的语义丰富化描述及研究设计指纹知识的识别提供理论支撑。第二章，全面介绍关于科技论文内容知识表示方法模型及指纹识别方法，最后对现有研究进行评价与分析。第三章，在详细分析了研究设计指纹识别特征因素的基础上，设计了本书的总体研究框架，并提出三大关键问题。第四章，构建研究设计指纹概念模型，该概念模型不仅能够结构化、语义化地组织科技论文的研究设计指纹，更能科学地将科技论文转换成机器可计算与理解的智能文献载体，辅助科研用户快速阅读与掌握科研成果，同时也为开展科技论文大数据方案挖掘的实现提供逻辑结构与框架指导。第五章，基于研究设计指纹识别相关影响因素的详细分析，提出并构建包含基于标引语义知识库的指纹识别方法和基于多个规则模式的指纹识别方法，以及两阶段多规则混合模式的指纹识别算法模型。该模型打破了传统的侧重某一种实现知识抽取的方法，将基于标引语义知识库的指纹识别方法和基于多个规则模式的指纹识别方法进行混合使用，并与机器学习方法相结合，能够克服单一使用某一种方法带来的对研究领域知识组织体系（KOS）依赖性较强、对知识模板及规则集合要求全而且质量要求高等问题。另外，本章后一阶段的研究中进一步利用深度学习相关的人工智能算法模型开展了指纹识别方法的研究，并针对人工智能领域的科技论文进行了试验，取得了不错的结果。第六章，进行了研究算法与模型的有效性验证，将来自爱思唯尔（Elsevier）的数据挖掘（Data Mining）研究主题的部分科技论文全文作为实验数据集，利用本书设计的算法模型计算识别和抽取这些论文的研究设计指纹，把算法模型计算的结果同领域专家标注的结果进行对比分析，验证该研究设计指纹识别算法模型的有效性。实验评估结果表明，本书提出并设计实现的研究设计指纹识别方法模型能够有效地对科技论文全文的研究设计指纹进行自动识别与抽取，达到了辅助科研用户快速从海量科技论文中发现与挖掘研究背景、研究假设、研究方法、研究工具及研究结论等知识线索对象的目标。第七章，对本书提出的研

究问题进行了总结与展望，为后续工作的展开提供了客观的参考背景。

本书在编写过程中得到恩师张晓林教授的悉心指导，很多的内容与研究思路是经过与恩师多次研讨、修改而确定的。在此感谢恩师张晓林教授的栽培，是他把笔者带入科学研究的道路，是他的谆谆教诲让笔者对科研有了新的认识，时刻提醒笔者要敢于提出问题，发现“痛点”。同时，在本书的研究、编写及审校过程中，余丽、王茜两位博士研究生及庞娜等硕士研究生给予了帮助并提出了许多修改建议，在此，表示诚挚的谢意。

在后续的研究中，笔者将继续探索文本语义理解与深度学习等智能算法先进成果的有机结合方式，实现原始方法创新和关键技术突破，构建垂直领域的研究设计指纹知识网络，多视角、细粒度、层次化地揭示科技论文内容中蕴含的丰富语义信息，以进一步提升数据深度知识挖掘利用能力，提升智能计算的数据组织能力，提升情报分析的数据感知、发展趋势预测、领域研究热点追踪能力，并为文献情报领域下一代智能化、个性化的精准知识服务提供可能的智能计算方法参考。由于笔者对该领域的认知水平有限，书中不足之处在所难免，承蒙读者不吝告知，将不胜感激。

钱　力

2018 年 11 月 5 日于中国科学院文献情报中心

目　录

第一章 绪 论

第一节 背 景

科技论文作为科学技术发展的重要战略资源，记录着科学真理验证过程、实验观测结果及研究结论等研究知识线索。科技论文中所涉及的研究设计（包括研究问题、研究方法、研究流程、研究工具、相关方法与技术参数设定等），为后续研究者提供了宝贵的方法论和研究操作基础，成为科研人员项目设计、研究方法有效性评估、研究过程问题诊断、研究结果鉴别与评价的重要基础。科研人员希望能够有工具来有效地回答“有谁用什么方法来解决这个问题”“哪些方法及其技术与参数设定能够更好地解决这个问题”等问题。但是，在科研文献数量迅速增加的环境下，在项目策划、设计、申请、立项、实施细节规划、实施管理等各个阶段，研究人员需要能够及时、准确地发现针对研究问题的各类研究设计及其细节，要系统地比较同一问题上不同的研究设计及其成效，要利用已有的各类研究设计及其执行效果来优化或调整自己的设计及其研究过程，要提供支持相应研究方法及其细节设置的知识证据链，而目前以主题词为主的 Data Mining 或者聚焦于文摘层面的知识发现技术还难以有效地完成这些任务。因此，设计并实现一

套自动识别与抽取论文研究设计的理论与技术方法体系就变得十分必要与迫切。

本书以上述背景与研究问题为切入点，面向科学研究过程，在分析科技论文文献模型与相关技术方法的基础上，参照规范的科学研究方法，基于科技论文构成的研究设计核心要素——研究设计指纹，设计并构建研究设计指纹概念模型、研究设计指纹自动识别模型和研究设计指纹自动识别计算方法体系。本书所提出与实现的研究设计指纹概念模型、研究设计指纹自动识别模型与计算方法体系不仅能够提高科技论文的结构化、内容语义化、机器可计算性及支持大数据挖掘的能力，更重要的是，能够实现对研究假设、研究背景、研究目标、研究方法、研究数据、研究工具、研究结果、研究结论及后续研究趋势等九种类型的研究设计指纹的快速发现，并在此基础上实现对研究设计这种针对研究问题的解决方案进行系统化挖掘发现（即科研问题解决方案挖掘）。

第二节　相关概念界定

一、研究设计指纹的内涵

目前，对研究设计指纹的内涵尚没有一个明确界定，根据本书要解决的实际问题，依据科技论文中涉及的研究方法之特征，将研究设计指纹定义为在一篇科技论文中能够唯一表示与描述科学研究设计的各个研究阶段与研究实体的重要知识单元。依据斯韦尔斯（Swales）、迈尔斯（Myers）及海兰（Hyland）等多位学者提出的科技论文写作意图共识，即论文研究内容是对当前研究领域做出的既有贡献，参照曼（Mann）关于科技论文内容的修辞结构理论（Rhetorical Structure Theory），借鉴侯赛因（Hossein）的《研究文献写作指南与要求》、巴尔达萨

雷（Baldassarre）的《写作与出版科技论文的指南》，以及《科技论文写作指南》，本书将研究设计指纹具体分为九种指纹特征类型，即研究假设、研究目标、研究背景、研究方法、研究数据、研究工具、研究结果、研究结论及研究趋势，界定了研究设计指纹的四个主要特征：①知识唯一性，即在遵守科研道德规范的前提下，这些重要知识单元所具有的研究设计指纹特征是唯一的，其特征的核心构成维度有作者与文章标题；②研究思维性，即研究设计指纹可以精炼地揭示一个科学研究设计的整体设计思路；③知识结构性，即研究设计指纹可以结构化地描述科学研究方法、过程和结果，将其中的重要知识进行抽取、组织与关联；④骨干网络性，即一篇科技论文利用研究设计指纹可以类似于网络骨干图一样，可视化地描绘科学研究中的骨干知识。

二、研究设计指纹类型

科技论文是科研过程的文字描述载体，依据研究设计指纹在各个科研过程中科研工作流的功能作用、所要表述的科技论文写作意图，以及在科技论文传播与共享利用过程中所起到的作用，可将研究设计指纹分为四个类型，即基础指纹、技术指纹、结论指纹和未来指纹。

（一）基础指纹

描述与揭示一个具体研究已有的知识基础，如研究背景、在什么样的应用场景下、哪些人在哪些研究方向上采取什么方法等。因此，基础指纹包括三种指纹，即研究背景、研究假设和研究目标。

（二）技术指纹

描述与揭示一个具体研究采用的研究方法等技术方法。因此，技术指纹包括

三种指纹，即研究方法、研究数据与研究工具。

（三）结论指纹

描述与揭示一个具体研究的实施结果是什么样的。因此，结论指纹包括两种指纹，即研究结果和研究结论。

（四）未来指纹

描述与揭示一个具体研究的下一步研究计划、研究方向等。因此，未来指纹包括一种指纹，即研究趋势。

第三节　研究目标与意义

一、研究目标

本书研究的总体目标是立足面向科研过程的思维，基于科技论文全文内容，构建一套行之有效的研究设计指纹识别方法模型体系，以辅助科学研究人员从海量科技论文资源中快速了解到所关注领域已经使用的研究方法、研究工具、研究结果等研究设计指纹，掌握当前研究进展状况，为科研人员进行卓越研究设计提供客观的科学依据。

本书的具体研究目标如下。

（1）提出并构建研究设计指纹概念模型。该模型可以动态、方便、快捷地将一篇或多篇科技论文描述成结构化、语义化、可计算机处理的研究设计知识要素，

以支持后期智能化的大数据挖掘与知识发现。

（2）设计并构建研究设计指纹识别模型。该模型利用影响科技论文全文内容识别研究设计指纹的主要因素，以技术方法为支撑，通过自然语言处理（Natural Language Processing，NLP）及机器学习等相关的信息技术，实现自动从科技论文全文中挖掘出描述科研问题解决方案的研究设计指纹，辅助用户快速从海量科技论文全文中发现使用的研究方法、研究工具等知识单元，为其创造出卓越设计提供新的技术路径支撑。

二、研究意义

（1）实现从科技论文全文中挖掘出科研问题解决方案，为科学研究课题的策划、优化、立项与论证提供技术方法工具。帮助科研人员快速发现研究主题相关的研究方法、研究结果、研究工具及相应的证据链信息，为科研人员更科学、合理地创造出卓越研究设计提供客观的科学依据与理论支撑。

（2）研究设计指纹的提出为描述科技论文提供了一种新的视角。传统的科技论文主要以文本段落进行组织，现代技术可以支持计算机可读的结构化描述（如 XML 格式），新型技术还可支持基于主题地图、主题关系等的文本内容描述。本书从“研究设计指纹”的视角，描述与表示一篇科技论文的细粒度知识。因此，可以精炼地揭示科学研究的设计思路，结构化地描述科学研究的方法、过程和结果，以及可视化地描绘科学研究的骨干知识网络。

（3）自动、快速地创建研究主题、研究领域的小型研究活动知识组织体系（Research Activity KOS）或领域研究活动词表，表明研究内容与研究方法的复杂关系。

（4）促进科技论文阅读范式由以人工阅读的方式向以机器辅助阅读的方式转变。

（5）研究设计指纹概念模型的构建为科技论文的智能化描述提供技术框架，也对未来的语义出版起到一定的示范引领作用。

（6）面向当前科学研究领域从数据资源范式向数据密集型范式的快速转变，本书设计与构建的技术方法体系为创造一种支撑上述范式转变的新型学术交流模式提供了更多的可能性。

第四节　研究思路和研究方法

一、研究思路

本书基于问题驱动的研究模式，从需求问题出发，调研与分析已有研究基础及可利用的研究方法，从提供服务引擎与创造应用场景两个方面，提出本书研究问题的解决方案，详见图 1.1。

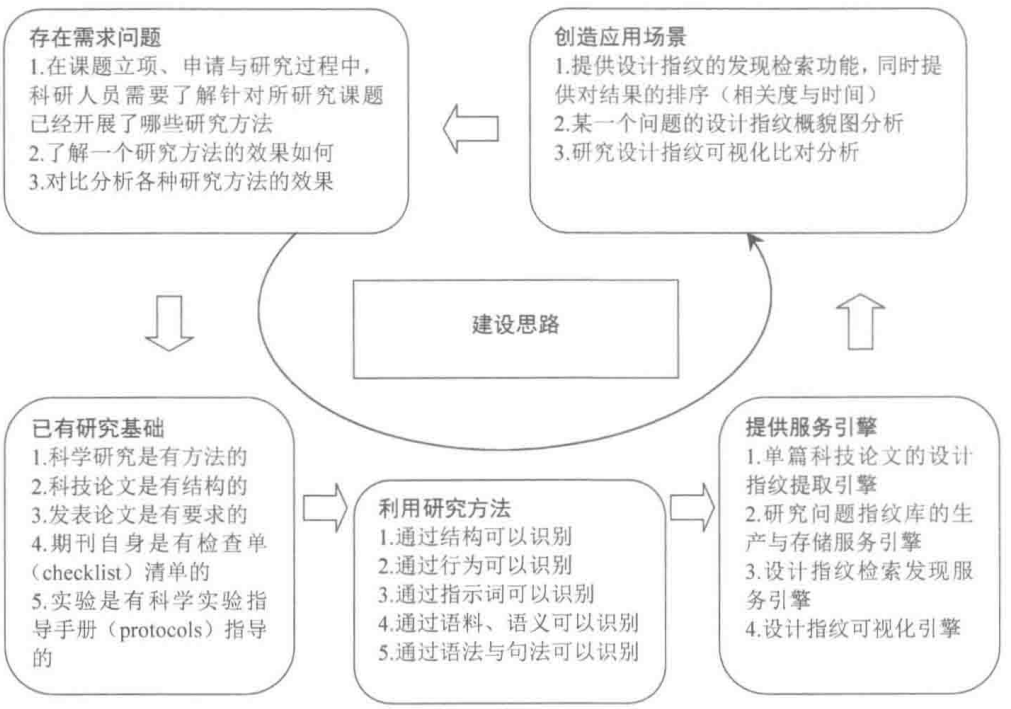


图 1.1　基于科技论文的研究设计指纹识别方法研究思路

二、研究方法

（一）系统调查法

全面了解科技论文的内容特征、结构特征与修辞特征，调研研究设计指纹识别的相关研究方向与方法，为本书的问题解决与技术方案的实施提供扎实的理论基础与技术指导。

（二）观察分析法

对实验型科技论文的全文内容、发表期刊的检查单规程、实验室计划的科学实验指导手册及科技论文撰写的手册与指南进行观察与分析，发现与掌握其中具有重要价值的知识信息点、规则，为指纹识别的计算方法体系提供算法设计基础。

（三）实验验证法

选定 Data Mining 研究主题的科技论文全文作为实验数据，使用论文设计与实现研究设计指纹识别模型和识别方法对实验数据进行指纹识别，通过对实验结果的全面分析，实现对指纹识别方法的有效性验证与科学评估。

第五节 研 究 内 容

一、理论基础框架研究

基于在研究所调研过程发现的现实需求问题，提出研究设计指纹概念，并以

此为切入点研究与分析研究设计指纹识别的相关理论模型和实现技术方法，提出并构建研究设计指纹概念模型，为将科技论文全文由传统的非结构化文本向智能化表示的转换提供一套标准体系框架。

二、研究设计指纹识别方法研究

从科技论文内容及结构等多个视角，全面分析影响研究设计指纹特征类型识别与判断的重要因素，在此基础上提出研究设计指纹识别总体技术框架，从指纹线索规则的发现、指纹识别模型的构建及指纹识别方法的设计与实现等多个方面，构建基于科技论文全文的两阶段与多规则结合的研究设计指纹识别技术方法体系，为构建微型知识组织体系、发现与追踪研究课题的已有研究方法及描绘单篇科技论文知识骨干网络提供工具平台支撑。

三、具体案例实证分析研究

以 Data Mining 研究主题的科技论文全文作为案例实证分析的实验数据，全面解析实验数据材料准备、研究设计指纹识别系统设计与实现及实证开展的具体应用等关键问题，结合实验结果对论文设计与提出的“基于科技论文的研究设计指标识别方法模型”的使用效果及性能进行分析与评价。