图书馆资源组织语义化理论及方法研究

刘　耀◎著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

国家社会科学基金项目“图书馆资源组织语义化理论及方法研究”（12BTQ006）资助

图书馆资源组织语义化理论及方法研究

刘　耀　著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS
·北京·

图书在版编目（CIP）数据

图书馆资源组织语义化理论及方法研究 / 刘耀著. —北京：科学技术文献出版社，2018.2
ISBN 978-7-5189-3642-7

Ⅰ.①图… Ⅱ.①刘… Ⅲ.①图书馆工作—文献资源—馆藏管理—研究
Ⅳ.① G253.5

中国版本图书馆 CIP 数据核字（2017）第 289068 号

图书馆资源组织语义化理论及方法研究

策划编辑：周国臻　责任编辑：周国臻　白建刚　责任校对：文　浩　责任出版：张志平

出　版　者　科学技术文献出版社
地　　址　北京市复兴路15号　邮编 100038
编　务　部　（010）58882938，58882087（传真）
发　行　部　（010）58882868，58882874（传真）
邮　购　部　（010）58882873
官方网址　www.stdp.com.cn
发　行　者　科学技术文献出版社发行　全国各地新华书店经销
印　刷　者　北京教图印刷有限公司
版　　次　2018 年 2 月第 1 版　2018 年 2 月第 1 次印刷
开　　本　710×1000　1/16
字　　数　195千
印　　张　12.75
书　　号　ISBN 978-7-5189-3642-7
定　　价　58.00元

前　言

光阴荏苒，我在中国科学技术信息研究所工作已有10年光景。2007年，我从北京大学计算语言学研究所博士后出站，之后一直从事自然语言处理、知识组织与知识工程相关的工作。多年在一线进行研究与实践工作，一路走来积累了丰富的经验，在不断迭代和升华中，逐渐形成了一套完整的覆盖资源获取、资源加工到知识服务全流程的思想理论。在该理论思想的指导下，围绕着技术目标，我领导团队逐步开展工程实践研究，对理论方法进行验证。同时，根据工程实践中的具体问题，进行关键技术攻关，对其中的一些流程、步骤、细节等进行了微调和补充，不断升级与完善PYROIS系统。PYROIS系统初建于2007年，2011年年初，PYROIS系统1.0版本上线，目前该系统3.0版本已研发完成。

我在很早之前就有想要写这本书的想法，分享研究实践中的一些想法和体会，为从事相关研究的人员提供一些思路、方法和参考，但一直没有时间。其间，也陆陆续续地发表了一些研究论文，但没有进行系统性阐述。PYROIS系统3.0版上线之后，空闲时间较以前也多了些，我便抓紧时间完成了早已有的写书的想法。

本书的主要思想是利用自然语言处理技术和机器学习方法对已有的公认领域知识进行重构并加以利用；在构建领域本体的基础上，对相关文献进行语义标注；并在大量内容相对语义化的基础上，结合传统组织资源，通过机器学习等方法，生成初始语义元数据；然后，在辅助平台的帮助下，实现图书馆资源组织语义

化过程与语义元数据体系构建同步进行，将语义标注文献与语义索引分别存储，实现图书馆资源组织语义化。

在思想理论指导及技术目标的驱动下，相关研究及实践工作体现在以下几个方面。

知识元数据库构建技术。从充分开发和利用百科全书资源的角度出发，利用自然语言处理技术和语言分析工具分析百科全书资源，发现其中隐含的知识点及其之间的内在关联，将大量的、不断出现的知识点结构化地组织和关联起来，构成智能连接的网状图，建立知识元数据库。

语义元数据构建与标注一体化技术。将图书馆资源组织语义化看作图书馆资源语义标注的浅层标注，是内容语义化与形式语义化交互实现的过程，基于 NLP 理论与方法，结合传统图书馆资源组织方式与领域本体构建技术，搭建辅助构建及标注一体化平台，实现语义元数据体系构建与资源组织语义化过程同步实施。

领域本体自动构建技术。通过领域本体构建、语义标注等信息技术的融合与集成，将行业领域知识合理分类，构建以知识点为基本单元的图书、期刊、专利数据库，以智能检索、知识语义导航、可视化等方式为科学研究、技术开发、工程设计、工程应用的开展提供知识服务，实现知识的共享与重用。

一体化爬虫技术。针对语义爬虫存在的不足，实现从一个概念出发，自动生成语义结构，用该语义结构指导爬虫抓取数据资源；同时，在爬虫爬行过程中，不断从数据库中抽取相关的概念及其之间的关系，并填充进语义结构，实现语义结构进化和爬虫爬行迭代一体化。

资源融合精准解析技术。以科技信息资源和服务体系为基础，

对专利、图书、论文、科技报告等资源进行精准解析与深度融合，突破海量知识数据的存储、组织、索引和检索的效率与性能，构建领域知识关联网络，以检索、分析、评价及可视化的方式提供服务，为政府决策、企业创新提供信息支撑。

受限领域语义资源构建技术。在领域资源匮乏的情况下，根据已知线索从多种渠道获取和利用外部资源，进行结构化、知识化处理，最大限度地完成资源内部知识体系的构建与资源的语义化标注处理，为行业领域知识服务应用场景提供数据支撑。

技术寻源技术。连接全球的创新企业、机构、专家、个人用户，汇集各方需求，实时匹配一流的技术解决方案，从传统的“靠自己”的研发模式转型为用户、资源全流程交互的开放研发模式，帮助用户获得产品的优先体验权和资源获得优先供货权，并助力小微企业快速实现产品迭代升级。

现在这个阶段主要围绕基础技术、行业共性技术及部分个性技术进行研发，并对思想理论进行验证。后续将吸收与结合大数据、深度学习等一些先进理论、方法和技术，进一步深入开展相关研究，以面向行业领域提供一站式的知识服务综合解决方案。

本书是国家社会科学基金项目“图书馆资源组织语义化理论及方法研究”（项目编号：12BTQ006）的最终研究成果。中国科学技术信息研究所的领导及科学技术文献出版社的编辑对本书的出版给予了大力支持，在此深表感谢。

在本书的编写过程中，作者参阅了大量的专业图书和文献，汲取了很多精髓，特别是引用了部分图表、数据等，在此，向有关作者表示诚挚的感谢。我的研究生郭志军、郑德举、张子渊、龚幸伟、王睿佳、孙丽君、帅远华、王明程、肖铮等人不同程度

地参与了前期的研究工作，研究生黄毅在资料整理方面做了大量的工作，在此一并表示感谢。同时，也向开发团队人员表示感谢。

由于水平和经验有限，加之书中涉及内容广泛，难免有不足之处，恳请同行专家和读者批评指正，以便在再版时进一步修订完善。

刘　耀

2018 年 2 月

目　录

图表目录