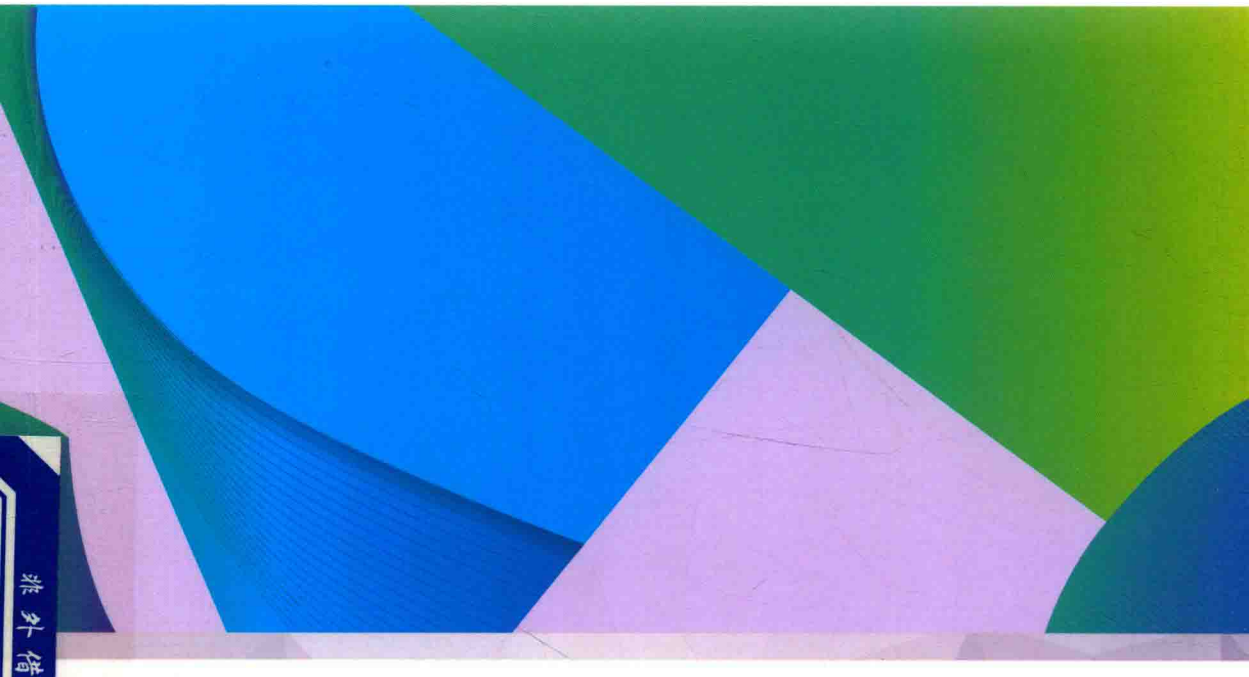
术语计算与知识组织研究

宋培彦 ◎ 著



非外借

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

术语计算与知识组织研究

宋培彦 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

术语计算与知识组织研究 / 宋培彦著. —北京：科学技术文献出版社，2018.6
ISBN 978-7-5189-4559-7

Ⅰ.①术… Ⅱ.①宋… Ⅲ.①自然语言处理—研究 Ⅳ.①TP391.1

中国版本图书馆 CIP 数据核字（2018）第 131143 号

术语计算与知识组织研究

策划编辑：周国臻　　责任编辑：张　红　　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) 58882868，58882870（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 6 月第 1 版　2018 年 6 月第 1 次印刷
开　　本　710×1000　1/16
字　　数　172千
印　　张　13
书　　号　ISBN 978-7-5189-4559-7
定　　价　58.00元

版权所有　违法必究
购买本社图书，凡字迹不清、缺页、倒页、脱页者，本社发行部负责调换

自序：走向智能化知识服务的“立心之战”

美国对中国芯片技术等领域进行“卡脖子”式的封锁，“缺芯之痛”引起举国关注。然而，看不见、摸不着但却同样重要的知识库，其作用并不逊色于芯片硬件，对智能信息处理和智能化知识服务这两个方面同时具有全局性、根本性和长期性的影响，堪称是不可或缺、不可替代、价值不可估量的“软芯片”。

智能信息处理需要强有力的知识库。这是因为，知识库是计算机智能信息处理的基石，也是未来智能计算机内嵌的核心技术，所以“不可或缺”；知识库本身是一个庞大的知识密集型系统，加上中国语言和文化自身的特殊性，因此很难指望外国人代劳或者从国外购买、坐享其成，而只能主要依靠国人自力更生、自主创新，因此“不可替代”；正是由于其重要性高、技术难度大、研发周期长而市场价值极为显著，因此其价值“不可估量”。离开了知识库，计算机还是擅长数据计算的“超级机器”，但绝不能称为一目千行、勤学善思的“电脑才子”，也与“智能”二字无缘。知识库是衡量“智能”水平高低的标志之一。

智能化知识服务也同样离不开知识库的有力支撑。在大数据环境下，图书情报界正在从传统的图书情报服务转向智能化知识服务，以术语库、叙词表、本体等为代表的知识组织工具本质上是作为计算机内置使用的知识库，扮演了计算机“大脑”的关键角色。智能检索、关联数据、语义网、知识图谱等智能化知识服务理念，都有赖于知识组织工具（知识库）。研制汉语知识库特别是专业领域知识库，可以为智能推送、精准检索、自动推理等智能化知识服务提供强有力的知识基础。可以说，知

识库建设涵盖了图书情报学（知识组织）、人工智能（自然语言处理）、语言学（术语学）、统计学乃至认知科学等多学科，堪称中国图书情报界开展智能化知识服务的“立心之战”，其科学意义和应用价值不言而喻。

所幸，我们有许多默默无闻、潜心研究的先哲时贤早就深刻意识到了这一点，凭借专业技能和满腔热忱肩负起为电脑“立心”、坚持自主创新的重任，数十年如一日，创立了知网 Hownet、概念层次网络 HNC、《同义词词林》、《汉语主题词表》等许多奠基性、开创性的知识工程。一些富有战略远见的创新型企业，如百度、阿里巴巴、腾讯等 IT 企业及出版业界，也紧紧抓住机遇、投入巨资，纷纷将知识图谱、网络百科、智能检索等基础研究变成落地生根的产品，服务于社会公众并取得良好经济效益。特别是，我国重点研发计划等重大工程着眼长远，将中文信息处理、类脑计算、知识工程等列为国家重大战略项目予以稳定支持，中文知识库这一“软芯片”研究与产业化基本与国际先进水平保持齐头并进，不仅成功避免了被外国技术封锁的被动局面，而且在某些领域已经走进世界舞台中央，从“跟跑”变成“领跑”。本书就是在这样的时代背景下应运而生的，最终希望对我国专业领域智能信息处理和知识服务水平的提升有所裨益。

同时，也应该注意到，我国在知识库方面的研究还存在一些亟待解决的问题。例如，传统中文知识库建设大多依靠专家手工劳动，不同程度地存在着效率偏低、成本高昂、更新迟缓、应用场景不明确等问题；或者过于强调统计模型、深度学习等技术手段，追求“短平快”，造成知识库准确性不足，难以支撑实际应用。因此，本书试图在知识组织的理论框架下，将术语计算技术引入知识库构建中，实现术语计算技术与知识组织理论的结合，探索更加精准化、智能化、开放性的知识库建设与服务新模式。这种结合不仅在理论上是可行的，而且对工程实践和实际应用也是有益的。

此为试读，需要完整PDF请访问：www.ertongbook.com

知识组织为术语计算提供了较完善的知识表示框架和数据基础。在大数据环境下，以术语为中心，将各类知识组织工具进行融合，形成统一的知识库，进而支撑专业知识的组织、挖掘、管理和服务，不仅可以“复用”叙词表、本体等现有的规范化知识体系，实现知识的继承和发展，而且能够实现对专业知识进行多维度、细粒度的组织、管理和利用，有助于消除散乱的术语造成的“信息孤岛”，并缓解“信息过载”，为用户提供精准的深层次知识服务，提高智能化知识服务的质量和效率。

术语计算为实现知识组织提供了有效技术手段。术语计算（Terminology Computing）是采用自然语言处理技术，对术语所包含的专业知识进行自动挖掘，以较小的颗粒度和可组合性对知识进行表示、推理和计算。通过对术语的挖掘和描述，构建细粒度、动态更新的术语知识库，有助于实现知识的动态聚合，进而以可视化、个性化的方式为用户提供知识服务。本书以自动分类、聚类、词义计算等术语计算技术为依托，形成快速、动态的知识发现与更新方法，从微观层面揭示知识的语义关系，以可视化方式直观展现知识关联，进而为构建词间关系、建立知识关联提供辅助手段，促进大数据时代知识动态更新和精准服务。

实践也充分证明，术语计算与知识组织在智能化知识服务这一根本目标上“殊途同归”，具有很强的可结合性，是构建新型知识库的有效方法。术语计算具有很强的操作性和应用性，在知识组织的框架中才能更好地发挥其应有的作用；现有的知识组织工具需要借助术语计算技术在更大范围内进行有序化组织和关联，才能充分发挥知识组织的价值、满足用户的知识需求。

万丈高楼平地起。以下项目提供的数据资源和应用场景，为本书提供了宝贵的实验基础，并验证了术语计算与知识组织结合的可行性。

（1）《汉语主题词表（工程技术卷）》编制与科技术语库建设。2010—2014 年，笔者作为核心研究人员参加了《汉语主题词表（工程技术卷）》的编制与修订工作。具有完全知识产权的新版《汉语主题词表

(工程技术卷)》核心词(术语)数量达到20余万条;科技术语库收录科技名词术语近400万词,并在语义计算、自动分类、自动标引等方面进行了积极探索,具有较为坚实的数据基础。

(2)英文超级科技词表与术语库建设。2011—2015年,在国家“十二五”科技支撑计划项目“面向外文文献的知识组织研究与示范应用”支持下,笔者作为术语库负责人,采集了约20万个工程技术领域核心概念、300多万个英文术语,形成了跨语言的知识组织工具,并用于国家科技文献中心NSTL外文文献的标引与检索服务,产生了良好的社会效益。

(3)术语服务系统建设。2011—2014年,笔者主持了国家社科基金项目“基于知识组织的术语服务研究”,参与研制了术语服务系统、主题自动标引系统等,具备中英文翻译、关系、属性、定义、分类信息等,并实现了对科技术语的语义链接和可视化,在国内一些专业机构开展示范应用。

(4)“国家科技管理信息系统”科技大数据中心建设。2015年以来,围绕“国家科技管理信息系统”建设,笔者主要从事科技大数据中心的设计与建设工作,实现与全国47个省市(含计划单列市、副省级城市、新疆生产建设兵团)的互联互通与共享服务,将国家科研项目、专家、成果等各类科技类数据进行深度组织和管理。术语计算和知识组织对于科技大数据中心建设发挥了重要作用。

(5)专家推荐系统建设。笔者还主持了2016年度国家社科基金项目“基于知识组织的科研项目评审专家发现研究”,通过知识组织工具的优化实现专家自动推荐,为科研项目评审、决策咨询等提供良好的支持。

上述研究经历让笔者更加深切地体会到,术语计算与知识组织有天然的紧密联系,具有重要研究价值。本书主要采用实证研究的方法,将术语理论、计算方法、模型与实验等尽可能与叙词表等知识组织工具进

行有效结合和验证，尽量使计算结果具有可重复性，以支撑知识组织的实际应用。同时，以知识组织理论框架为纲，将散乱的术语融合成有机整体，并以智能检索、术语服务、数字出版、科技辅助决策等作为应用场景进行验证，也有助于体现术语计算的效果和价值。理论、方法、技术与应用，在这里相得益彰、并行不悖，构成本书的主线。

本书命名为《术语计算与知识组织研究》，蕴含的目标是希望促进术语计算与知识组织的跨学科结合，寻求智能化知识服务的新突破。笔者主要研究方向是知识组织和自然语言处理，恰好具备图书情报学、计算机和语言学等方面复合型知识结构，加上在知识组织工具建设、术语计算、术语服务等方面具备较好的实际工作经验和科研条件，包括主持国家社科基金项目2项、发表学术论文20余篇，获得2项发明专利、6项软件著作权等，这些成果均围绕术语计算与知识组织进行，并在一些重要的科技管理信息系统中产生了较好的应用效果，种种机缘巧合，得以管窥其中奥妙。当然，这是一条鲜花与荆棘相伴、机遇与挑战并存、幸福与艰辛同在的羊肠小道，我们不能奢望“毕其功于一役”，试图通过一本书或项目解决所有问题。倘若能够吸引更多的有志者携手前行，开辟更加宽广的科学大道，那将是本书最大的成功。

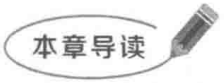
宋代理学家张载说“为天地立心，为生民立命，为往圣继绝学，为万世开太平”，这是做学问的高层次境界，千百年来激励人心、催人奋进。在大数据和人工智能时代构建高水平知识库，这既是为我国智能信息处理事业“立心”“立命”，有助于提高我国的智能化技术水平和服务能力，也是科技工作者“继绝学”“开太平”的责任担当和价值体现。我们应该与时俱进，在术语计算和知识组织的结合上有更多、更大的作为，推动更强劲的“中国芯”率先巍然屹立在中华大地上。

目　录

此为试读，需要完整PDF请访问：www.ertongbook.com

图表目录

第一章　绪论：知识为体，术语为用



本章主要以术语和知识组织的关系为对象，介绍以下内容。

- 术语与知识组织的关系是什么？二者应如何结合？
- 大数据时代，知识组织的现状、趋势和现实需求是什么？为什么术语计算变得更加重要？
- 本书的研究目的和相关背景是什么？如何阅读和使用本书？

1　术语：打开知识宝库的一把钥匙

1.1　术语与知识组织：如影随形的黄金搭档

术语（Terminology）是“在特定专业领域中一般概念的词语指称”[1]，是科学共同体记录科研成果、进行知识交流的重要交际工具。随着知识的不断积累，术语的数量和所涉及的领域也在逐渐丰富，已经成为人们开展科技交流、学术研究和生产劳动的基础性工具。近年来，随着自然语言处理技术的深入发展和海量文献的迅猛增长，术语学与知识工程、自然语言处理技术等深度结合，人们开始以术语为基础，从工程化的角度进行专业领域知识的表示和挖掘，形成可供计算机利用的专业领域知识库，这已经成为术语研究的重要课题，并引起人工智能领域的关注[2]。

知识组织（Knowledge Organization）是“以知识为对象的诸如整理、加工、表示、控制等一系列组织化过程及其方法”[3]，本质上对知识进

此为试读，需要完整PDF请访问：www.ertongbook.com