HZ BOOKS

大 数 据 管 理 丛 书

短文本数据理解

王仲远 编著

机械工业出版社
China Machine Press

大/数/据/管/理/丛/书

短文本数据理解

王仲远　编著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

短文本数据理解 / 王仲远编著 . —北京：机械工业出版社，2017.2
（大数据管理丛书）

ISBN 978-7-111-55881-1

I. 短…　II. 王…　III. 文本编辑　IV. TP311.11

中国版本图书馆 CIP 数据核字（2017）第 013386 号

本书围绕短文本理解的各项需求及挑战，创造性地提出了概念化模型作为短文本理解的核心技术，可以广泛应用于搜索引擎、广告系统、智能助手等场景中，是大数据管理不可或缺的部分，具有较高的实际应用价值。 全书深入浅出、案例丰富，适合知识图谱、自然语言处理、信息检索、人工智能等方向的研究生阅读，也能为相关领域研究人员和开发人员提供重要参考。

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：余　洁　　责任校对：李秋荣
印　　刷：北京文昌阁彩色印刷有限责任公司　　版　　次：2017 年 5 月第 1 版第 1 次印刷
开　　本：170mm × 242mm　1/16　　印　　张：10
书　　号：ISBN 978-7-111-55881-1　　定　　价：69.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com

‖ 丛书前言

当下大数据技术发展变化日新月异，大数据应用已经遍及工业和社会生活的方方面面，原有的数据管理理论体系与大数据产业应用之间的差距日益加大，而工业界对于大数据人才的需求却急剧增加。大数据专业人才的培养是新一轮科技较量的基础，高等院校承担着大数据人才培养的重任。因此大数据相关课程将逐渐成为国内高校计算机相关专业的重要课程。但纵观大数据人才培养课程体系尚不尽如人意，多是已有课程的“冷拼盘”，顶多是加点“调料”，原材料没有新鲜感。现阶段无论多么新多么好的人才培养计划，都只能在 20 世纪六七十年代编写的计算机知识体系上施教，无法把当下大数据带给我们的新思维、新知识传导给学生。

为此我们意识到，缺少基础性工作和原始积累，就难以培养符合工业界需要的大数据复合型和交叉型人才。因此急需在思维和理念方面进行转变，为现有的课程和知识体系按大数据应用需求进行延展和补充，加入新的可以因材施教的知识模块。我们肩负着大数据时代知识更新的使命，每一位学者都有责任和义务去为此“增砖添瓦”。

在此背景下，我们策划和组织了这套大数据管理丛书，希望能够培

养数据思维的理念，对原有数据管理知识体系进行完善和补充，面向新的技术热点，提出新的知识体系/知识点，拉近教材体系与大数据应用的距离，为受教者应对现代技术带来的大数据领域的新问题和挑战，扫除障碍。我们相信，假以时日，这些著作汇溪成河，必将对未来大数据人才培养起到“基石”的作用。

丛书定位：面向新形势下的大数据技术发展对人才培养提出的挑战，旨在为学术研究和人才培养提供可供参考的“基石”。虽然是一些不起眼的“砖头瓦块”，但可以为大数据人才培养积累可用的新模块(新素材)，弥补原有知识体系与应用问题之前的鸿沟，力图为现有的数据管理知识查漏补缺，聚少成多，最终形成适应大数据技术发展和人才培养的知识体系和教材基础。

丛书特点：丛书借鉴 Morgan & Claypool Publishers 出版的 Synthesis Lectures on Data Management，特色在于选题新颖，短小精湛。选题新颖即面向技术热点，弥补现有知识体系的漏洞和不足(或延伸或补充)，内容涵盖大数据管理的理论、方法、技术等诸多方面。短小精湛则不求系统性和完备性，但每本书要自成知识体系，重在阐述基本问题和方法，并辅以例题说明，便于施教。

丛书组织：丛书采用国际学术出版通行的主编负责制，为此特邀中国人民大学孟小峰教授(email：xfmeng@ruc. edu. cn)担任丛书主编，负责丛书的整体规划和选题。责任编辑为机械工业出版社华章分社姚蕾编辑(email：yaolei@hzbook. com)。

当今数据洪流席卷全球，而中国正在努力从数据大国走向数据强国，大数据时代的知识更新和人才培养刻不容缓，虽然我们的力量有限，但聚少成多，积小致巨。因此，我们在设计本套丛书封面的时候，特意选择了清代苏州籍宫廷画家徐扬描绘苏州风物的巨幅长卷画作《姑苏繁华图》(原名《盛世滋生图》)作为底图以表达我们的美好愿景，每

本书选取这幅巨卷的一部分，一步步见证和记录数据管理领域的学者在学术研究和工程应用中的探索和实践，最终形成适应大数据技术发展和人才培养的知识图谱，共同谱写出我们这个大数据时代的盛世华章。

在此期望有志于大数据人才培养并具有丰富理论和实践经验的学者和专业人员能够加入到这套书的编写工作中来，共同为中国大数据研究和人才培养贡献自己的智慧和力量，共筑属于我们自己的“时代记忆”。欢迎读者对我们的出版工作提出宝贵意见和建议。

大数据管理丛书

主编：孟小峰

大数据管理概论

孟小峰　编著

2017 年 5 月

异构信息网络挖掘：原理和方法

[美]孙艺洲(Yizhou Sun)　韩家炜(Jiawei Han)　著

段磊　朱敏　唐常杰　译

2017 年 5 月

大规模元搜索引擎技术

[美]孟卫一(Weiyi Meng)　於德(Clement T. Yu)　著

朱亮　译

2017 年 5 月

大数据集成

[美]董欣(Xin Luna Dong)　戴夫士·斯里瓦斯塔瓦(Divesh Sriva-stava)　著

王秋月　杜治娟　王硕　译

2017 年 5 月

短文本数据理解

王仲远　编著

2017 年 5 月

个人数据管理

李玉坤　孟小峰　编著

2017 年 5 月

位置大数据隐私管理

潘晓　霍峥　孟小峰　编著

2017 年 5 月

移动数据挖掘

连德富　张富峥　王英子　袁晶　谢幸　编著

2017 年 5 月

云数据管理：挑战与机遇

[美]迪卫艾肯特·阿格拉沃尔(Divyakant Agrawal)　苏迪皮托·达斯(Sudipto Das)　阿姆鲁·埃尔·阿巴迪(Amr El Abbadi)　著

马友忠　孟小峰　译

2017 年 5 月

推荐序一

短文本理解是伴随着搜索引擎、社交网络及聊天机器人等应用场景而兴起的一个研究课题。它是近些年的一个研究热点，且对未来人工智能的发展有重要的影响。由于短文本字词少、歧义大、不遵守语法规则等特点，传统自然语言处理技术如句法分析器等难以直接应用于短文本。因此，研究人员不得不另辟蹊径来解决机器理解短文本的问题。

从2009年起，我在微软亚洲研究院领导一个小组从事短文本的研究工作。2010年7月，本书作者王仲远加入微软亚洲研究院并参与这方面的研究。我们及组里其他同事共同开发了一个Web规模的知识库系统Probase，尝试解决知识尤其是常识的获取、表示及应用问题。我们认为“概念”对于理解短文本的语义至关重要，正如纽约大学著名心理学教授Gregory L. Murphy在其代表性著作《The Big Book of Concepts》中提到“Concepts are the glue that holds our mental world together”(概念是我们思想的粘合剂)。通过Probase，我们尝试着将一些心理学研究的课题可计算化，并取得了很大的成果。2011年，仲远开始在中国人民大学攻读在职博士生，我很荣幸又成为他的博士生导师。之后，仲远在围绕Probase的工程项目、学术研究中不断突飞猛进，取得了一个又一个成果。

2013年，我离开微软，仲远接手了Probase项目。他不断深化基于

Probase 所构建的短文本理解概念化模型，并获得了国际著名学术会议 ICDE 2015 最佳论文奖。在 2016 年的国际自然语言处理学术会议 ACL 上，仲远和我共同作了一个报告“Understanding Short Texts”。我们将短文本理解的方法简要分为隐性模型和显性模型两大类。隐性模型主要是基于词向量和深度神经网络的模型，其主要缺点是模型为一个“黑盒子”，结果常常难以具体化解释。而另一方面，显性模型主要依赖于知识库系统或语义网络，其可解释性强于隐性模型，但知识的获取及表示是一大挑战。尤其是知识质量与覆盖率，更是会直接影响显性模型的最终效果。

我非常高兴地看到仲远将这些年的研究成果整理成书。这本书对短文本概念化问题进行了详细的介绍，既有单实体概念化模型，也有短文本概念化模型，并介绍了概念化模型的一些典型应用。全书结构合理，系统性强，并且本书许多章节都包含了大量实例与插图，便于读者理解背后的技术模型，也使得本书有很强的实用性和阅读性。

希望本书能为知识图谱、自然语言处理、信息检索、人工智能等相关领域研究人员和开发人员提供重要参考。我愿全力推荐本书给广大读者。

Haixun Wang

Facebook Research Scientist & Engineering Manager

2016 年 9 月 26 日于美国 Palo Alto

‖ 推荐序二

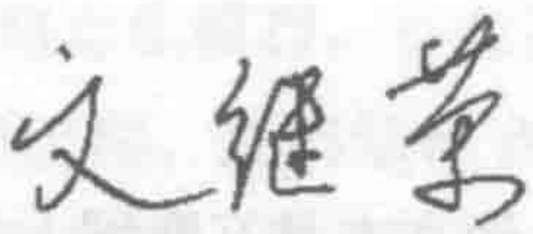
短文本是互联网上广泛存在的一种文本数据，如搜索引擎查询、广告及推荐系统关键词、社交网络聊天记录、产品的用户评论等。然而，由于短文本“短”的特性，使得机器理解其语义面临极大的挑战。以英文搜索引擎的查询为例，97%的搜索查询所包含的词数少于或等于 8 个，其中更是有 63%的搜索查询只包含一两个词。因此对于短文本，机器必须从极为有限的上下文中，尝试挖掘出丰富而有效的信息，这是关乎机器人工智能的基础性研究，对许多实际应用场景具有至关重要的意义。

本书围绕短文本理解的各项需求及挑战，创造性地提出了概念化模型作为短文本理解的核心技术，为解决机器短文本理解这一问题迈出了重要的一步。本书涵盖了如下创新性研究内容：1)提出了基于概率的属性提取与推导，并挖掘了动词、形容词等非实体词与概念之间的语义关联，为短文本理解奠定了基础，完善了短文本理解所需的语义网络；2)针对短文本理解的概念化模型，通过解决短文本中单实体和多实体的概念化问题，克服了短文本较稀疏、噪声多、歧义大的特点，将短文本转为机器可以计算的一种显性概念向量表示方法，这成为短文本理解的一种新的解决方案；3)针对短文本中的主题词与修饰词检测问题，提出了一种基于概念化、面向开放领域的无监督检测机制。

本书作者王仲远是我的博士生，也曾是微软亚洲研究院最年轻的主

管研究员之一。他在微软亚洲研究院工作以及博士研究生就读期间在顶级学术会议和期刊上发表了一系列与短文本相关的论文，并在提炼和系统化这些工作的基础上写就了其博士论文。作为其导师，我很欣慰地看到他不辞辛苦地将其博士论文整理成册，将其中的理论和技术介绍给更多的读者，从而推动国内相关研究领域的发展。

全书结构清晰，深入浅出，以大量实例来解释其背后的技术难点与解决方案，并展示了在实际广告系统中的应用实例。相信本书对广大的科研工作者、研究生及从事相关工作的算法工程师都具有重要的参考价值。我向广大读者大力推荐这本书籍！



国家“千人计划”特聘专家，中国人民大学信息学院院长

2016 年 9 月 26 日

前 言

当今世界，每天都有数十亿的短文本产生，比如搜索查询、广告关键字、标签、微博、问答、聊天记录等。与长文本(如文档)不同，短文本具有如下特性：首先，短文本通常不遵守语法规则；其次，短文本由于字数少，本身所包含的信息也较少。前者使得传统的自然语言处理方法不能直接适用于短文本，而后者则意味着短文本理解不得不依赖于外部信息。简而言之，短文本具有较稀疏、噪声大、歧义多的特点，因而机器理解短文本面临极大的挑战。

而另一方面，随着近些年人工智能技术的重大突破，尤其是大规模知识图谱以及深度学习技术的出现，使得机器理解短文本出现新的曙光。研究者们提出了许多将文本转换成机器所能理解的内部表示方法。这些方法可以分为三类：1)隐性知识表示方法，如基于深度学习产生的向量表示法；2)半显性知识表示方法，如主题模型；3)显性知识表示方法，如概念化模型。这些方法各有优缺点。一般而言，前两类方法适用广泛，已有若干成熟应用，但其所产生的模型难以被人类理解，因此优化较为困难。而后一类方法正蓬勃发展，涌现出许多新的模型，并已在许多大型互联网公司如 Google、微软内部使用。如果读者对这几类方法的概况有进一步了解的兴趣，可以参见本书作者在国际自然语言处理顶级学术会议 ACL 2016 上的一个专题教程(Tutorial)报告“Understanding Short

Texts”(理解短文本)(主页地址：http：//www. wangzhongyuan. com/tutorial/ACL2016/Understanding-Short-Texts/)。

本书主要介绍基于知识图谱进行显性短文本理解的方法，即由笔者提出的创新性概念化模型，并对不同情况下的概念化过程进行深入分析与探讨。本书许多章节的内容依托于发表在国际相关领域顶级学术会议或期刊上的技术论文，并已实际应用于微软的众多产品中(如必应搜索、广告系统、MSN 查询推荐、Office 365 等)。

尤为值得一提的是，笔者在微软亚洲研究院领导开发多年的大型知识库系统 Probase 也于近期由微软研究院正式发布。发布的正式名称为“Microsoft Concept Graph”(微软概念图谱)，网址为 https：//concept. research. microsoft. com/。有兴趣的读者可以访问该发布网址以获得更多详细信息，本书许多章节中的模型都是构建在这个概念图谱之上(书中称其为知识库、语义网络或 Probase)。读者也可以从该发布网址中获得微软从海量互联网网页中所挖掘出的知识图谱数据，以便作进一步研究使用。

本书的内容和组织结构

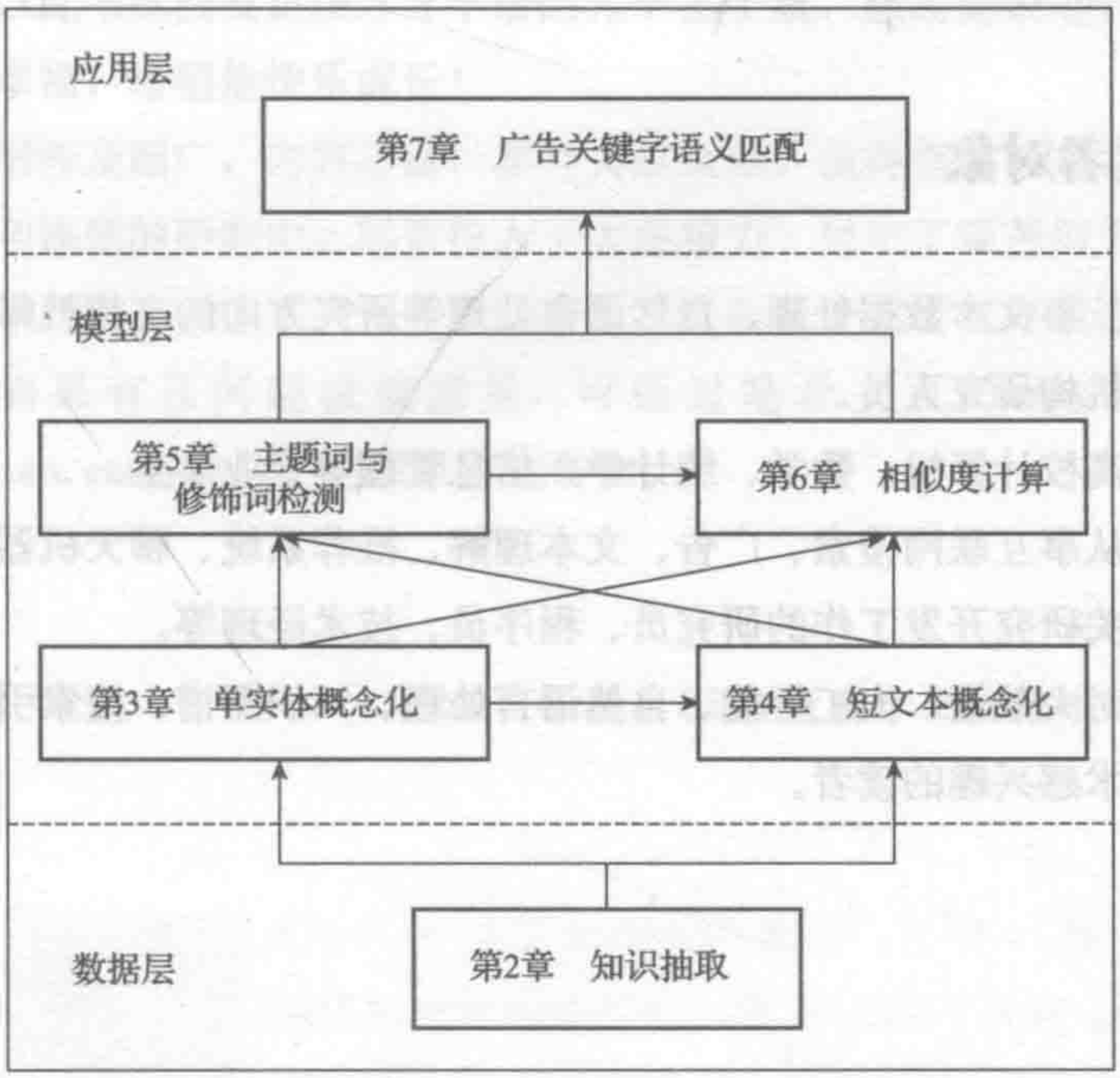
本书内容依照数据层、模型层和应用层逐步展开介绍。其中，第 2 章为数据层，第 3～6 章为模型层，第 7 章为应用层。

本书组织结构如下：

第 1 章为“短文本理解及其应用”。主要介绍短文本理解的研究背景及意义，分析短文本理解的研究现状。

第 2 章为“基于概率的属性提取与推导”。主要介绍一种在语义网络层，为百万级的概念推导出属性的方法。

第 3 章为“单实体概念化模型”。介绍了一种基于典型性和点互信息(PMI)将单实体映射到概念空间的基本层次概念化(Basic-level Conceptualization，BLC)方法。



第 4 章为“基于概念化的短文本理解”。介绍一种基于概念化的查询理解方法，把短文本(如搜索引擎中的查询关键字)所包含的实体映射到概念空间上，从而支持机器进行进一步的计算。

第 5 章为“基于概念化的短文本主题词与修饰词检测”。基于概念化模型，将大量实体级别的“主题词-修饰词”对映射为精细且精确的带权重的概念模式，进而进行主题词与修饰词的检测。

第 6 章为“基于概念化的词相似度计算”。利用概念化模型，将词映射为一种语义表示，从而计算任意两个词之间的语义相似度值。

第 7 章为“基于概念化的海量竞价关键字匹配”。展示了本书所介绍的模型在实际系统中的应用，把短文本概念化成一组相关概念，通过测量它们在概率空间的相似度，对于给定的查询选择相关的竞价关键字。

第 8 章为“短文本理解研究展望”。指出了短文本理解方向未来的研究工作。

本书读者对象

- 从事文本数据处理、自然语言处理等研究方向的高校教师及科研机构研究人员。
- 高校计算机、数学、统计学、信息管理等专业学生。
- 从事互联网搜索、广告、文本理解、推荐系统、聊天机器人等相关研究开发工作的研究员、程序员、技术经理等。
- 对大数据、人工智能、自然语言处理、知识图谱、搜索引擎等技术感兴趣的读者。

致谢

本书内容凝结了笔者在微软亚洲研究院多年研究成果的结晶。在此衷心感谢我的导师孟小峰教授、文继荣教授、王海勋博士将我带入了学术的殿堂。在他们的指导下，我从一名普通的高校学生成长为一名合格的研究员，并且能在一些研究领域得到同行的认可。感谢我在微软亚洲研究院的同事李红松、宋阳秋、邵斌、宋睿华、窦志成、闫峻、纪蕾、马维英等，他们在我的研究中给予了热心帮助，与他们的讨论也对我的研究思路有很大的启发。感谢复旦大学肖仰华副教授、北京大学邹磊副教授、上海交通大学朱其立教授，与他们共同合作论文是一种荣幸。感谢在微软亚洲研究院实习过的李培培、Taesung Lee、王芳、胡志睿、华雯、赵可君、程健鹏、张大卫、郝泽慧、徐昊文、王鹏伟、李英杰等四十余位实习生，与他们一起讨论、工作，才有一个个将创新想法变为现实的可能。感谢胡莎、韩家龙同学，他们的睿智、热情、友善、诚恳时刻影响着我。感谢家人一直以来对我的支持。感谢我的妻子、我的父母、我的姐姐，他们的理解、支持与鼓励是我一步步前行的动力。感谢所有还未提及的老师、同学和朋友们！

谨以此书献给我正在牙牙学语的儿子王子航，感谢他带给我的无尽欢乐与幸福，希望他快乐成长！

本书涉及面广，内容丰富，参考文献众多。值得指出的是，在全书的撰写和课题的研究中，尽管投入了大量精力、付出了艰苦努力，但受知识水平所限，书中不当之处在所难免，诚恳希望读者批评指正并不吝赐教。如果有任何建议或意见，可通过笔者主页（http：//wang-zhongyuan. com/en/）上的联系方式告知。

王仲远

2016 年 9 月 25 日凌晨于北京西绦胡同

作者简介 ‖

王仲远　博士，美国 Facebook 公司 Research Scientist。加入 Facebook 前，他是微软亚洲研究院的主管研究员，领导微软研究院的两个知识图谱项目 Probase(即微软的概念知识图谱/Microsoft Concept Graph)和 Enterprise Dictionary(企业知识图谱项目)，以及一个人工智能助手项目 Digtal Me。他多年来专注于知识图谱及其在文本理解方面的研究，已在 SIGMOD、VLDB、ICDE、IJCAI、AAAI、CIKM、EMNLP 等国际顶级学术会议上发表论文 30 余篇，其中包括 ICDE 2015 最佳论文奖。他也是国际自然语言顶级学术会议 ACL 2016 Tutorial "Understanding Short Texts"的主讲人之一。目前已出版技术专著 2 本，拥有美国专利 5 项。他的研究兴趣包括：文本理解、知识库系统、自然语言处理、深度学习、数据挖掘等。