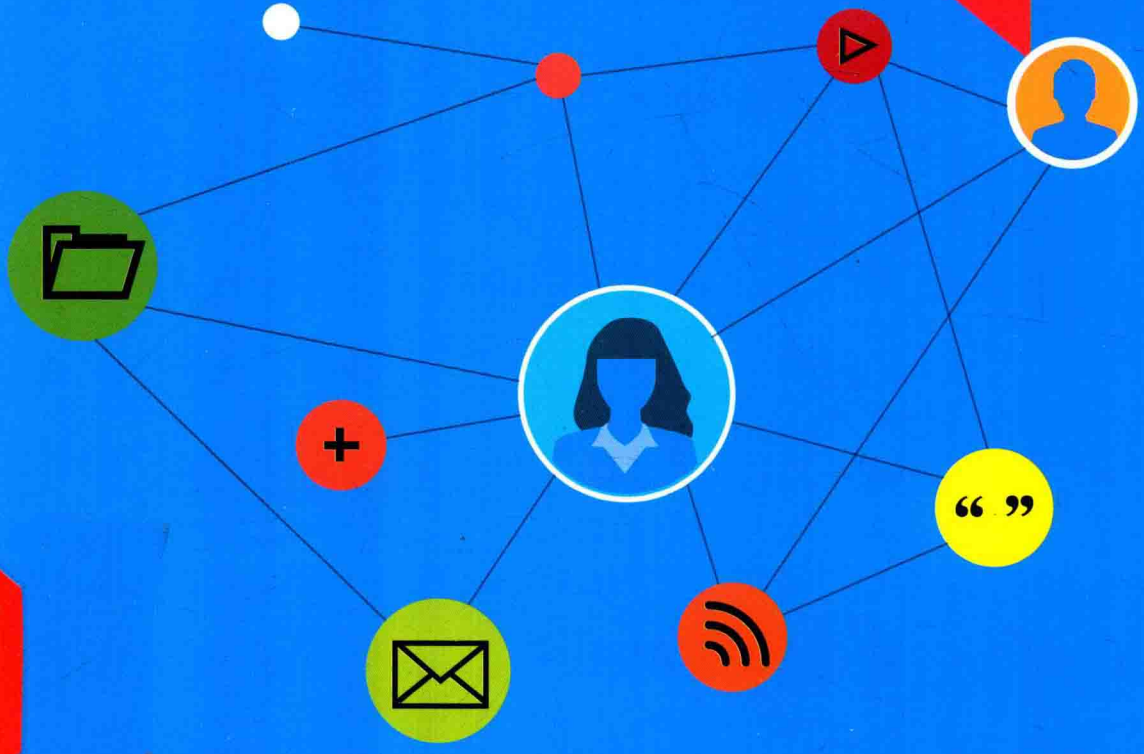
开放网络知识计算

——模型、方法与应用

王元卓　贾岩涛
林海伦　程学旗　著



清华大学出版社

开放网络知识计算

——模型、方法与应用

王元卓　贾岩涛
林海伦　程学旗　著

清华大学出版社
北京

内 容 简 介

网络大数据是指“人、机、物”三元世界在网络空间(cyberspace)中彼此交互与融合所产生并在互联网上可获得的大数据,简称网络数据。本书提出了开放知识网络的概念,以概率论、图论、矩阵分析、组合优化等为模型基础,给出了一套从开放知识的感知与获取、开放知识的融合与更新、开放知识的推断与预测,到开放知识计算引擎的构建及系统应用的开放知识处理流程。深入探讨了开放知识网络的建模与计算方法,并通过开放网络知识库和应用系统,介绍了典型应用案例,全面、系统地展示了本领域最新的研究成果和进展。

本书可作为计算机、通信、信息等相关专业的教师、研究生和大学高年级学生的教材或教学参考书,也可供行业大数据分析、商业情报挖掘、语义检索、知识问答等方面的研究人员和工程技术人员参考。

图书在版编目(CIP)数据

开放网络知识计算:模型、方法与应用/王元卓等著. —北京:清华大学出版社,2018
ISBN 978-7-302-49143-9

Ⅰ. ①开…　Ⅱ. ①王…　Ⅲ. ①计算机网络—网络计算　Ⅳ. ①TP393.027

中国版本图书馆 CIP 数据核字(2017)第 322753 号

责任编辑:薛　慧
封面设计:何凤霞
责任校对:赵丽敏
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 装 者:北京泽宇印刷有限公司
经　　销:全国新华书店
开　　本:185mm×260mm　**印　　张**:22.25　**插　　页**:2　**字　　数**:547 千字
版　　次:2018 年 5 月第 1 版　　**印　　次**:2018 年 5 月第1次印刷
定　　价:69.00 元

产品编号:074763-01

前言

网络大数据是指“人、机、物”三元世界在网络空间(cyberspace)彼此交互与融合所产生并在互联网上可获得的大数据,简称网络数据。当前,网络大数据在规模与复杂度上的快速增长对现有 IT 架构的处理和计算能力提出了挑战。网络大数据包含大量有价值的信息,根据其产生方式的不同可以分为 Web 内容数据、Web 结构数据、自媒体数据和日志数据等。

这些有价值的信息往往会通过某些属性或关系互相关联,这些反映相互关系的链接整合起来就是一个网络。这个网络中的数据具有多源异构、交互性、时效性、社会性、突发性和高噪声等特点,不但非结构化数据多,而且数据的实时性强。数据自身的信息、数据间的关联信息以及网络的结构特征等都隐藏在这样的数据网络中,网络大数据往往以复杂关联的数据网络这样一种独特的形式存在。有效利用网络大数据的主要任务不是获取越来越多的数据,而是对数据去冗分类、去粗取精,从数据中挖掘知识,对大数据网络后面的知识进行深入分析。

面对复杂关联、动态变化、来源多样的网络知识,建构开放网络知识的有效模型是一个重要基础,它应该支持对开放网络知识自适应的感知、增量的更新、自动或半自动的新知识抽取等,并具有较为完备的基础理论支撑。同时,从应用角度,开放网络知识计算需要建立一套算子体系,实现面向领域的开放网络知识库的快速构建,并更好地支持知识演化、多源知识融合、隐性知识推断和时序知识预测等一系列问题的解决。

本书主要以概率论、图论、矩阵分析、组合优化等为模型基础,深入探讨了开放知识网络的建模与计算方法,并通过开放网络知识库和应用系统,介绍了典型应用案例。

本书共 14 章,从结构上可分为 4 个部分。

第一部分主要介绍研究对象,包括第 1 章网络大数据和第 2 章开放网络知识。其中,在第 1 章网络大数据中,介绍了网络大数据研究体系,从网络空间感知与数据表示、网络大数据存储与管理体系、网络数据挖掘和社会化计算以及网络大数据平台系统与应用 4 个方面回顾了相关领域的新近发展,探讨了网络大数据研究方向和

所面临的挑战，并展望了网络大数据的主要研究方向。在第 2 章开放网络知识中，对当前国内外知名的开放网络知识库以及其支持的应用进行了分析和论述，并从开放网络知识库的构建以及基于开放网络知识库对信息检索与数据挖掘方面的应用方法和技术现状进行了综述，最后，展望了开放网络知识计算的应用和发展方向。

第二部分主要介绍开放网络知识计算的模型研究所需的基础理论和针对开放网络知识计算提出的模型方法。包括第 3 章概率论基础、第 4 章图论基础、第 5 章矩阵分析基础、第 6 章组合优化知识等基础理论知识。在此基础上，第 7 章给出了开放知识网络模型的表示方法、理论分析与证明。

第三部分介绍基于开放知识网络模型的知识计算方法，包括开放知识网络的构建、知识的融合与更新、知识推断和知识预测等知识计算的全生命周期。包括第 8 章～第 11 章。其中，第 8 章开放网络知识库的构建方法，包括开放文本中的领域概念抽取、实体属性抽取、实体关系抽取、领域概念的细化等；第 9 章从实体融合、关系融合、类别融合、自适应更新 4 个方面介绍知识融合与更新方法；第 10 章介绍的知识推断方法包括静态关系的推断和非时序动态关系的推断两个方面；第 11 章解决知识的预测问题，即给出时序的动态关系的知识推断方法。

第四部分介绍开放网络知识计算系统和应用场景。包括第 12 章～第 14 章。其中，第 12 章介绍现有的知识库与知识分析系统，包括早年由 Metaweb 公司创建的 Freebase 知识库、德国马普研究所的 Yago 知识库、微软公司的 Probase 知识库、谷歌公司的基于 Knowledge Graph 的知识计算系统、谷歌公司的基于 Knowledge Vault 的知识计算系统、大数据公司 Palantir 的知识计算系统、卡耐基-梅隆大学的 NELL 系统以及国内知名知识库和知识系统。针对现有的知识库构建技术缺乏有效的评价标准的问题，最后，给出了开放网络知识库构建的多维指标体系与量化评价方法。第 13 章将给出开放知识计算引擎，即 OpenKN 的整体架构与特点，以应对网络大数据下知识计算的实际需求。第 14 章将从人物谱系关系画像与分析、领域事件的演化态势分析、新闻语义推荐 3 个方面，分别探讨开放知识网络的应用场景与实际应用效果。

本书可供计算机、通信、信息等相关专业的教师、研究生和大学高年级学生作为教材或教学参考书，也适合大数据分析、商业情报挖掘、语义检索、知识问答等方面的研究人员和工程技术人员阅读使用。

本书涉及的研究工作得到了国家自然科学基金项目(No. 61572469，No. 61402442，No. 61602467)和国家重点基础研究发展计划“973”项目(No. 2014CB340400)和北京市自然科学基金项目 (No. 4154086)等的资助，在此表示深深的谢意！

中国科学院计算技术研究所的硕士研究生赵泽亚、李晓静、常雨骁、陈新蕾、蔡朋杉、李曼玲、仇韫琦、苏家林等人参与了本书的材料收集、撰写和排版等工作，在此一并表示感谢。

由于作者水平所限，加之开放知识计算方法的研究和应用仍处于不断发展和变化之中，书中错误和不足之处在所难免，恳请读者予以指正。

作　者

2017 年 5 月

目录

第一部分　网络大数据中的开放知识

第二部分 模 型 理 论

第三部分 计 算 方 法

第一部分

网络大数据中的开放知识

“人、机、物”三元世界融合的网络空间(cyberspace)中的网络大数据存在数据规模巨大、数据关联复杂、数据状态演变等显著特征。其规模和复杂度的增长远远超出了符合摩尔定律增长的机器处理和计算能力。网络大数据带来了宝贵的机遇,同时也存在着巨大挑战。本书的第一部分包括第1章网络大数据和第2章开放网络知识。

在第1章网络大数据中,介绍了网络大数据研究体系,从网络空间感知与数据表示、网络大数据存储与管理体系、网络数据挖掘和社会化计算以及网络大数据平台系统与应用4个方面回顾了相关领域的新近发展,探讨了网络大数据的研究方向和所面临的挑战,并展望了网络大数据的主要研究方向。在第2章开放网络知识中,对当前国内外知名的开放网络知识库及其支持的应用进行了分析和论述,并从开放网络知识库的构建以及基于开放网络知识库对信息检索与数据挖掘方面的应用方法和技术现状进行了综述,最后,展望了开放知识网络及其应用的未来发展方向。

第 1 章
网络大数据

近年来，随着互联网、物联网、云计算、三网融合等 IT 与通信技术的迅猛发展，数据的快速增长成为许多行业共同面对的严峻挑战和宝贵机遇，可以说信息社会已经进入了大数据(big data)时代。大数据的涌现不仅改变着人们的生活与工作方式、企业的运作模式，甚至还引起科学研究模式的根本性改变。

1.1 网络大数据

一般意义上，大数据是指无法在一定时间内用常规机器和软/硬件工具对其进行感知、获取、管理、处理和服务的数据集合[1]。网络大数据是指“人、机、物”三元世界在网络空间彼此交互与融合所产生并在互联网上可获得的大数据，简称网络数据。当前，网络大数据在规模与复杂度上的快速增长对现有 IT 架构的处理和计算能力提出了挑战。著名咨询公司 IDC 发布的研究报告指出，未来全球数据总量年增长率将维持在 50%左右，到 2020 年，全球数据总量将达到 40ZB($1Z=10^{21}$)。

网络大数据中包含大量有价值的信息，根据其产生方式的不同可以分为 Web 内容数据、Web 结构数据、自媒体数据、日志数据。其中，Web 内容数据主要是通过互联网网页产生和发布的数据，它既可以是文字、文本、消息，也可以是图片音视频等，以及 HTML、Java scripts、Interstitial 间隙窗口、Microsoft Netshow、Flash 等所产生或解析的数据。如今，Web 内容数据量呈指数级增长，例如检索网页的总量达 500 亿，在线图书网页达 7.5 亿，其中，英文维基百科数量达 427 万个页面，中文百科数据达 900 万个页面。Web 内容数据的特点既包括数据量巨大、内容信息丰富，还具有动态更新快，多源异构等特点。Web 结构数据是指 Web 页面间的结构数据，主要包括页面间的超链接关系和 Web 的组织结构。伴随着 Web 内容数据的增长，Web 页面间的链接关系也呈现出大规模增长的趋势。

自媒体数据主要是指通过以 Facebook、Twitter 等为代表的社交网络中产生的用户生成数据(user generated content, UGC)，具有空

前的规模性和群体性，数据总量巨大，数据变化非常快。1min 内，Twitter 上新发的数据量超过 10 万条；Facebook 用户每天分享的内容条目超过 25 亿个，数据库中的数据每天增加超过 500TB。此外，自媒体数据还具有十分复杂的内在关系，超过 10 亿的 Facebook 用户的好友关系和超过 5 亿的 Twitter 用户之间的关注关系构成了极为复杂的关系网络。

日志数据主要指各种网上服务提供商积累的系统和用户操作的日志记录，比如 Google、百度等搜索引擎提供商积累的用户搜索行为日志等。此类数据的特点是，具有大量的历史性数据，同时数据增速极快、数据访问吞吐量巨大。以 Google 为例，目前有超过 200 个谷歌文件系统 GFS(Google File System)集群在运行，而每个集群有 1000～5000 台机器，每个 GFS 都存储着高达 5PB 的数据；成千上万台机器需要的数据都从 GFS 集群中检索，这些集群中数据读写的吞吐量可高达 40GB/s，每天都在产生着富含大量知识的数据。IBM 将大数据的特点总结为 3 个 V，即大量化(volume)、多样化(variety)和快速化(velocity)。首先，网络空间中数据的体量不断扩大，数据集合的规模已经从 GB、TB 到了 PB，而网络大数据甚至以 EB 和 ZB 等单位来计数。IDC 的研究报告称，未来 10 年全球大数据将增加 50 倍，管理数据仓库的服务器的数量将增加 10 倍，以迎合 50 倍的大数据增长①。其次，网络大数据类型繁多，包括结构化数据、半结构化数据和非结构化数据。在现代互联网应用中，呈现出非结构化数据大幅增长的特点，至 2012 年末，非结构化数据占有比例达到互联网整个数据量的 75%以上。这些非结构化数据的产生往往伴随着社交网络、移动计算和传感器等新技术的不断涌现和应用。再次，网络大数据往往呈现出突发涌现等非线性状态演变现象，因此难以对其变化进行有效的评估和预测。另一方面，网络大数据常常以数据流的形式动态、快速地产生，具有很强的时效性，用户只有把握好对数据流的掌控才能充分利用这些数据。

近几年，网络大数据越来越显示出巨大的影响力，正在改变着人们的工作与生活。2012 年 11 月《时代》杂志撰文指出奥巴马总统连任成功背后的秘密，其中的关键是对过去两年来相关网络数据的搜集、分析和挖掘②。目前，eBay 的分析平台每天处理的数据量高达 100PB，超过了纳斯达克交易所每天的数据处理量。为了准确分析用户的购物行为，eBay 定义了超过 500 种类型的数据，对顾客的行为进行跟踪分析③。每年的互联网购物季，都发生着大规模的商业活动，其中，在“双十一”期间，天猫淘宝系网站的销售总额已经突破千亿元人民币。淘宝之所以能应对如此巨大的交易量和超高并发性的分析需求，得益于对往年的情况，特别是用户的消费习惯、搜索习惯以及浏览习惯等数据所进行的综合分析④。

网络大数据给学术界也同样带来了巨大的挑战和机遇。网络数据科学与技术作为信息科学、社会科学、网络科学、系统科学等相关领域交叉的新兴学科方向正逐步成为学术研究的新热点。近年来，“Nature”和“Science”等刊物相继出版专刊来探讨对大数据的研究。2008 年，“Nature”出版专刊“Big Data”，从互联网技术、网络经济学、超级计算、环境科学、生物医药等多个方面介绍了海量数据带来的挑战[2]。2011 年，“Science”推出关于数据处理的

① http://www.emc.com/

② http://swampland.time.com/

③ http://www.china-cloud.com/

④ http://server.51cto.com/

专刊“Dealing with data”，讨论了数据洪流(data deluge)所带来的机遇[3]。特别指出，倘若能够更有效地组织和使用这些数据，人们将得到更多的机会发挥科学技术对社会发展的巨大推动作用。

1.2　网络大数据研究的意义

总体而言，网络大数据研究的重要性体现在以下几个方面。

(1) 网络大数据对捍卫国家网络空间的数字主权、维护社会稳定、推动社会与经济可持续发展有着独特的作用。信息化时代，国家层面的竞争力将部分体现为一国拥有网络大数据的规模、活性以及对数据的解释与运用的能力。国家在网络空间的数字主权也将是继海、陆、空、天这 4 个空间之后另一个大国博弈的空间。在网络大数据领域的落后，意味着失守产业战略制高点，意味着国家安全将在网络空间出现漏洞。为此，2012 年 3 月，美国政府整合 6 个部门投资 2 亿美元启动“大数据研究和发展计划”。在该计划中，美国国家科学基金会提出要“形成一个包括数学、统计基础和计算机算法的独特学科”。该计划还强调，大数据技术事关美国的国家安全，影响科学研究的步伐，还将引发教育和学习的变革。这意味着网络大数据的主权已上升为国家意志，直接影响国家和社会的稳定，事关国家的战略安全。

(2) 网络大数据是国民经济核心产业信息化升级的重要推动力量。“人、机、物”三元世界的融合产生了大规模的数据，如何感知、测量、利用这些网络大数据成为国民经济中许多行业面临的共同难题，成为这些行业数字化、信息化的障碍和藩篱。如何使不同行业都能突破这一障碍，关键在于对网络大数据基本共性问题的解决。譬如，对于非结构化数据的统一表示与分析，目前缺少有效的方法和工具。因此，通过对网络大数据共性问题的分析和研究，使企业能够掌握网络大数据的处理能力或者能够承受网络大数据处理的成本与代价，进而使整个行业迈入数字化与信息化的新阶段。从这个意义上看，对网络大数据基础共性问题的解决将是新一代信息技术融合应用的新焦点，是信息产业持续高速增长的新引擎，也是行业用户提升竞争能力的新动力。

(3) 网络大数据在科学和技术上的突破，将可能诞生出数据服务、数据材料、数据制药等战略性新兴产业。网络数据科学与技术的突破意味着人们能够理清数据交互连接产生的复杂性，掌握数据冗余与缺失双重特征引起的不确定性，驾驭数据的高速增长与交叉互连引起的涌现性(emergence)[4]，进而能够根据实际需求从网络数据中挖掘出其所蕴含的信息、知识甚至是智慧，最终达到充分利用网络数据价值的目的。涌现性是指由低层次的多个元素构成高层次的系统时展示出的每个单一元素所不具备的性质。网络数据不再是产业环节上产生的副产品，相反地，网络数据已成为联系各个环节的关键纽带。通过对网络数据纽带的分析与掌握，可以降低行业成本、提升行业效率和生产力。因此，可以预见，在网络数据的驱动下，行业模式的革新将可能催生出数据材料、数据制造、数据能源、数据制药等一系列战略性的新兴产业。

(4) 大数据引起了学术界对科学研究方法论的重新审视，正在引发科学研究思维与方法的一场革命。科学研究最初只有实验科学，随后出现了理论科学，研究各种定律和定理。由于在许多问题上，理论分析方法变得太过复杂以至于难以解决难题，人们开始寻求模拟的方法，这又产生了计算科学。而大数据的出现催生了一种新的科研模式，即面对大数据，科