HZ BOOKS
华章IT

Mining Heterogeneous Information Networks
Principles and Methodologies

大数据管理丛书

异构信息网络挖掘
原理和方法

[美] 孙艺洲（Yizhou Sun） 著
韩家炜（Jiawei Han）
段磊 朱敏 唐常杰 译

机械工业出版社
China Machine Press

大/数/据/管/理/丛/书

Mining Heterogeneous Information Networks

Principles and Methodologies

异构信息网络挖掘

原理和方法

[美] 孙艺洲（Yizhou Sun） 著
韩家炜（Jiawei Han）
段磊 朱敏 唐常杰 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

异构信息网络挖掘：原理和方法 /（美）孙艺洲（Yizhou Sun），（美）韩家炜（Jiawei Han）著；段磊，朱敏，唐常杰译．—北京：机械工业出版社，2016.9
（大数据管理丛书）
书名原文：Mining Heterogeneous Information Networks: Principles and Methodologies

ISBN 978-7-111-54995-6

I. 异… II. ① 孙… ② 韩… ③ 段… ④ 朱… ⑤ 唐… III. 数据采集－研究 IV. TP274

中国版本图书馆 CIP 数据核字（2016）第 236589 号

本书版权登记号：图字：01-2016-3481

Authorized translation from the English language edition, entitled Mining Heterogeneous Information Networks: Principles and Methodologies, 9781608458806 by Yizhou Sun, Jiawei Han, published by Morgan & Claypool Publishers, Inc., copyright © 2012.

Chinese language edition published by China Machine Press, copyright © 2017.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage retrieval system, without permission from Morgan & Claypool Publishers, Inc. and China Machine Press.

本书中文简体字版由美国摩根 & 克莱普尔出版公司授权机械工业出版社独家出版。未经出版者预先书面许可，不得以任何方式复制或抄袭本书的任何部分。

本书介绍了异构信息网络挖掘的原理和方法，包括基于排名的聚类与分类、基于元路径的相似性搜索和挖掘、关系强度感知挖掘，以及若干有前景的研究方向。本书是伊利诺伊大学香槟分校数据挖掘高级课程的参考教材，适合作为数据挖掘方向的研究生教材，也适合数据挖掘研究人员和专业技术人员参考。

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：和　静　　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：11.25
书　　号：ISBN 978-7-111-54995-6　　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

版权所有 · 侵权必究
封底无防伪标均为盗版
本书法律顾问：北京大成律师事务所　韩光 / 邹晓东

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

此为试读，需要完整PDF请访问：www.ertongbook.com

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

‖ 译者序

作为大数据时代铺路石的数据采集技术，近年来的发展突飞猛进。它带来两个巨大变革：一方面，现实生活中各项事物间的联系愈发紧密，以社交网络为例，一个成熟的社交网络不仅包含着人与人之间的各种联系，还包含着人与时间、空间、机构等其他相关因素的联系；另一方面，我们也更关注于发掘多源数据所蕴含的丰富、复杂、有趣的未知知识。如何将这些现实世界中多源且异构的关系处理为计算机可表达并可计算的形式？异构信息网络理论及其技术成为解决这一问题的利器。

本书是一本极具学术价值的介绍异构信息网络相关概念及分析方法的专著，在介绍异构信息网络基本概念的基础上，结合实例讲述了异构信息网络中的聚类与分类、基于元路径的相似性搜索和关系预测、关系强度感知挖掘等重要内容。本书循序渐进、用例丰富，容易为读者理解，且内容新颖全面，涉及异构信息网络的基本概念、与其他类型网络的差别、异构信息网络分析及应用等。本书适合数据挖掘方向的高年级本科生、硕士生和博士生阅读，也适合相关研究和应用技术人员参考。

本书的作者孙艺洲博士致力于异构信息网络挖掘研究，曾荣获ACM SIGKDD 2013 博士论文奖，是一位冉冉升起的数据挖掘研究者新星。本书的另一位作者韩家炜教授在数据挖掘领域发表了很多具有广泛影响力的高水平著作和论文。本书就是两位作者关于异构信息网络挖掘

研究和应用的总结。

本书翻译工作主要由段磊、朱敏、唐常杰完成。四川大学研究生杨皓、刘璐、晏力、秦攀、高超、王文韬等对本书的翻译提供了不少帮助，译者谨在此对他们表示感谢，并向在翻译过程中给予我们大力支持的华章公司的姚蕾、朱劼、和静三位老师表示衷心的感谢。

译者在翻译过程中力求忠于原著，新的专业术语尽量符合原著语义。但由于水平和时间有限，译文难免有错误和不妥之处，恳请读者批评指正。

段　磊

2016 年 5 月于四川大学

现实世界中物理的和抽象的数据对象相互联系，形成巨大、交织的网络。通过将这些数据对象和它们间的交互结构化为多种类型，这样的网络变成半结构化异构信息网络。现实世界中，绝大多数处理大数据的应用，包括相互联系的社交媒体和社交网络，科学的、工程的或医学的信息系统，在线电子商务系统以及大量数据库系统，都可以被结构化为异构信息网络。因此，如何有效地分析大规模异构信息网络成为一个有趣而重要的挑战。

本书讲述了异构信息网络挖掘的原理和方法。与许多现有网络模型将相互连接的数据视作同构图或网络不同，半结构化异构信息网络模型充分利用网络中各类型节点和链接的丰富语义，并从网络中发现大量丰富知识。半结构化异构网络建模为挖掘相互联系的数据提供了一系列崭新原理和有力方法，包括：(1)基于排名的聚类与分类；(2)基于元路径的相似性搜索和挖掘；(3)关系强度感知挖掘，以及若干有潜力的进展。本书介绍了异构信息网络挖掘的前沿研究，并指出了若干有前景的研究方向。

关键词：信息网络挖掘；异构信息网络；链接分析；聚类；分类；排名；相似性搜索；关系预测；用户引导聚类；概率模型；真实应用；高效可伸缩算法

孙艺洲(Yizhou Sun) 2012 年在伊利诺伊大学香槟分校获得计算机科学博士学位。她目前是美国加州大学洛杉矶分校计算机科学系助理教授，主要研究兴趣集中在大规模信息和社交网络挖掘，研究领域为数据挖掘、数据库系统、统计学、机器学习、信息检索和网络科学，并对新奇问题的建模及大规模真实应用的可伸缩算法的开发较为关注，已有 60 余篇著作发表于专著、期刊和重要会议。基于异构信息网络挖掘方面的研究工作，她受邀在 EDBT 2009、SIGMOD 2010、KDD 2010、ICDE 2012、VLDB 2012 和 ASONAM 2012 等高水平会议上作主题报告。她曾获得 ACM KDD 2012 最佳学生论文奖，ACM SIGKDD 2013 博士论文奖，2013 年 Yahoo 学术职业发展奖，2015 年美国科学基金会职业成就奖(NSF CAREER Award)。

此为试读,需要完整PDF请访问: www.ertongbook.com

韩家炜(Jiawei Han) 伊利诺伊大学香槟分校的 Abel Bliss 教授，研究领域包括数据挖掘、信息网络分析、数据库系统和数据仓库。他是若干国际会议的主席或程序委员，包括 KDD、SDM 和 ICDM 等国际知名会议的程序委员会主席，VLDB 会议的美洲协调员。他创办了《ACM Transactions on Knowledge Discovery from Data》学报并任创始主编，在数据挖掘、数据库和信息网络领域发表论文 600 余篇。他是 ACM 和 IEEE Fellow，曾获得 2004 年度 ACM SIGKDD 创新奖，2015 年度 IEEE 计算机学会技术成就奖，2009 年度 IEEE 计算机学会 Wallace McDowell 奖和 2011 年度伊利诺伊大学香槟分校 Daniel C. Drucker 杰出教员奖。他的著作《数据挖掘：概念与技术》是全球范围内被广泛采用的教科书。

目 录

此为试读,需要完整PDF请访问: www.ertongbook.com