大 数 据 管 理 丛 书

实体识别技术

申德荣 寇月 聂铁铮 于戈 等编著

机械工业出版社
China Machine Press

大/数/据/管/理/丛/书

实体识别技术

申德荣 寇月 聂铁铮 于戈 编著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

实体识别技术 / 申德荣等编著 . —北京：机械工业出版社，2017.9
（大数据管理丛书）

ISBN 978-7-111-58161-1

I. 实… II. 申… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字（2017）第 243365 号

本书基于作者多年对数据集成的相关研究工作，从多角度阐述实体识别技术，主要包括相似度计算算法、实体识别的分块技术、典型的基于机器学习的实体识别技术和基于关系的实体记录识别技术，以及新型的实体识别技术（包括基于时间的实体识别技术、基于众包的实体识别、隐私保护下的实体识别）等内容。全书深入浅出、案例丰富，适合数据集成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：佘 洁 责任校对：李秋荣
印 刷：北京诚信伟业印刷有限公司 版 次：2017 年 10 月第 1 版第 1 次印刷
开 本：170mm×242mm 1/16 印 张：13.5
书 号：ISBN 978-7-111-58161-1 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

大数据、小数据、无数据：网络世界的数据学术

[美]克莉丝汀 L. 伯格曼(Christine L. Borgman)　著

孟小峰　张祎　赵尔平　译

2017 年 8 月

实体识别技术

申德荣　寇月　聂铁铮　于戈　等编著

2017 年 10 月

‖前　言

在大数据时代，数据驱动的数据分析与挖掘已成为各领域决策的客观依据。然而，由于不同的数据源有不同的描述实体的方式，并且可能存在拼写错误、缩写方式不同、描述格式不同、属性值缺失、属性值随着时间演化等特点，导致描述真实世界同一实体的不同数据对象存在差异。实体识别将一个或多个数据源中描述真实世界同一实体的数据对象识别出来，提升集成的大数据资源的质量。

实体识别最早出现在人口普查和医疗卫生等社会公共服务领域，很早就受到公共机构的重视和依赖，从而促进了实体识别的研究。实体识别已经有几十年的研究历史，出现了许多有效的实体识别技术。在大数据时代的今天，实体识别在多个领域有着广泛的应用需求，包括客户关系管理、人口普查、医疗卫生、网购比价、国家安全、引文数据库、垃圾邮件检测、关联的数据(Linked Data)、机器阅读等。

本书作者多年来一直从事数据集成相关研究，实体识别是提升数据集成质量的关键技术之一。在国家 973 计划、国家自然科学基金、国家 863 计划等课题的支持下，作者分别针对关系数据对象识别、复杂数据空间中的数据对象识别、具有时间特性的数据对象识别、隐私保护下的

数据对象识别等方面进行了深入研究。本书基于已有相关研究，综述了当前已有的实体识别技术，目的是为相关研究者提供一定借鉴作用。

本书共分八章，主要内容包括概述、相似度计算算法、实体识别的分块技术、典型的基于机器学习的实体识别技术和基于关系的实体识别技术，以及新型的实体识别技术(包括基于时间模型的实体识别、基于众包的实体识别、隐私保护下的实体识别)等。

本书由东北大学计算机科学与工程学院计算机科学系申德荣、寇月、聂铁铮、于戈等撰写。其中，申德荣、于戈负责本书前言部分，申德荣、于戈、孙琛琛负责第1章，韩姝敏、寇月负责第2章，聂铁铮负责第3章，孙琛琛、聂铁铮负责第4章，寇月负责第5章，申德荣、韩姝敏负责第6章，孙琛琛、韩姝敏负责第7章，申德荣、孙琛琛负责第8章。参加本书撰写的还有硕士研究生刘宏、汪潜等。全书由申德荣统稿，由于戈教授主审。

我们在撰写本书过程中，覆盖了经典实体识别技术和新型的实体识别技术，跟踪了该学科的新发展和新技术，力求本书具有先进性和实用性。但由于作者学识有限，一定存在许多不足之处，敬请专家和学者批评指正。

第1章

概　述

1.1　实体识别问题的提出

大数据时代，数据生成的速度和更新频率远超过去[1]，商业组织、公共部门和政府部门都在面临大量数据的冲击，高效地处理和分析这些数据有助于商业决策、公共政策制定、政府职能提升和国家安全维护。数据管理与数据挖掘是数据研究的核心领域。数据管理聚焦于高效地集成、存储和查询海量数据；数据挖掘则致力于从已有的数据中发掘潜在的信息和价值。

在大规模信息系统和大型的数据挖掘项目中，经常需要将来自多数据源的数据进行集成，提高数据质量，实现数据信息互补，为后续的数据分析与挖掘提供一个完整的、干净的、统一的数据集。集成后的数据集比之前分裂的多个数据集的价值更大，可以从中挖掘出更多的知识，为用户提供更多有价值的信息。在此过程中，一个非常重要的步骤是实体识别[2-13]，即将描述相同真实世界实体的不同数据对象识别出来，从而在数据融合时，能够将描述相同实体的数据对象合并成一个干净的、统一的、健全的数据记录，提高集成数据的质量。

实体识别的直接原因是数据冗余的存在。根据数据源是否单一，可以将数据冗余分为两类：单数据源数据冗余和跨数据源数据冗余。单数据源数据冗余通常由于在加入新的数据记录的时候没有执行严格的重复检测或者完全没有执行重复检测。比如，一个大型商场(在不同城市有分店)的客户信息记录，同一个客户可能进行了多次客户信息登记，而接待人员没有发现这些重复登记。造成这个状况的原因多种多样，如每次登记的姓名有差别、工作单位不同、家庭住址不同等。跨数据源的数据冗余则更加显而易见，当将多个数据集合成一个数据集时，来自不同数据集的数据记录很有可能描述相同的实体。比如，两家公司实行合并后，对它们的客户信息进行整合，需要将他们共同的客户信息找出来。跨数据源的实体识别中，模式匹配是前提。

实体识别中的数据对象(即数据记录)描述真实世界的实体，通常包括多个属性，如姓名、年龄和地址等。这里的数据对象是结构化的，符合一定的数据格式，比如客户信息的数据记录包括姓名属性、年龄属性、电话号码属性、地址属性和工作单位属性。实体识别中最常见的一类数据对象是描述人的数据对象，如商业数据库中的客户记录、公司数据库中的员工记录、航空公司数据库中的乘客记录、医院数据库中的病人记录和医疗保险记录、国家安全部门数据库中的嫌疑犯记录和政府数据库中的纳税人记录等[8]。除了人，还有其他的实体类型，如商业记录、出版记录、引文记录、产品记录等。例如，在商品比价应用中，由于不同电商网站的描述格式不同，识别出哪些商品记录描述着相同的商品有一定难度；还有引文记录中的会议(或出版社)全称和缩写的识别、作者单位全称与简称的识别等[8]。

1.2 实体识别研究的发展历史

实体识别起源于统计学家和公共健康研究领域，在单数据库内或多数据库中识别对应同一实体的重复记录。1946 年，Dunn 应用术语“记

录链接”(record linkage)[14]来描述现实世界中每一个个体的生命溯源，即从生到死整个生命周期中个体所经历的信息，如健康信息、社会保障、结婚、离婚等记录信息。20 世纪 50 年代末和 60 年代初，Howard Newcombe 等[15-16]提出应用计算机自动处理实体识别过程，并提出了基于概率的记录链接方法的成功理念。基于 Newcombe 的思想，在 1969 年，两个统计学家 Ivan Fellegi 和 Alan Sunter[17]为实体识别引入了正式的数学模型。

1999 年，由学者 Winkler[18]扩展并提高了最初的模型，最显著的工作是引入了字符串近似比较函数[19]来捕捉字符串的变化情况，以及应用期望(EM)算法[20]来改进概率记录链接中匹配参数的估计。同时，数据库研究团队从数据清洗需求出发，提出了重复记录识别技术[21]，用于改进数据库的质量[22]。但是，数据库研究者并没有采用由 Fellegi 和 Sunter 提出的基于概率的匹配方法，而是应用近似串比较函数计算属性相似度[23-24]，并通过属性比较来发现相似的记录[21,25]。

随着数据的丰富，计算机领域中有关实体识别的研究备受关注，尤其是在数据挖掘、机器学习和信息获取领域。此外，数据库和数据仓库研究团队[26]相应地也提出了一些新的实体识别技术[27]，如利用机器学习、自然语言处理和基于图的方法来改进数据质量。除此之外，近些年来还呈现出了面向时间记录的实体识别[28-29]，改善具有时间演化特性的同一实体的识别准确性；基于众包的实体识别[30-31]，通过混合人机来提升实体识别的准确性。同时，隐私保护下的实体识别[32,33]也成为了关注热点，以支持隐私数据的实体识别。

根据识别对象的数据源的种类划分，已有的实体识别工作主要包括：在关系数据库、Deep Web 数据库上的实体(记录)识别[34]；Web 上的实体识别[35-36]；语义 Web(RDF 数据)上的实体识别[37-38]；数据仓库中的实体识别[39-40]；非结构化文档中的实体识别[41]；复杂数据如 XML 数据、图数据、复杂网络上的实体识别[42]；社会网络中的实体识别[43-44]。