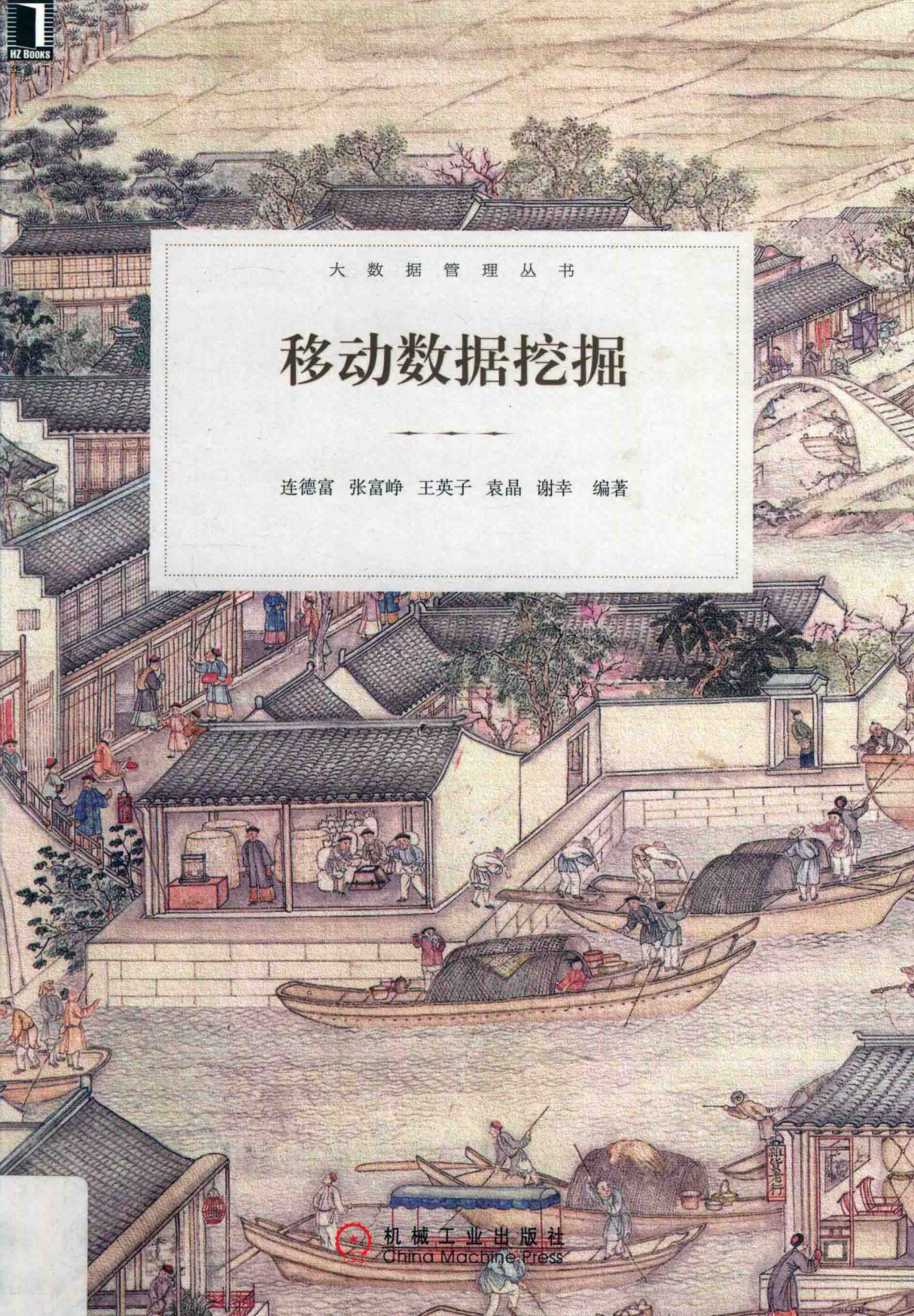


大 数 据 管 理 丛 书

移动数据挖掘

连德富 张富峥 王英子 袁晶 谢幸 编著

机械工业出版社
China Machine Press

大/数/据/管/理/丛/书

移动数据挖掘

连德富 张富峥 王英子 袁晶 谢幸 编著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

移动数据挖掘 / 连德富等编著 . —北京：机械工业出版社，2017.3
（大数据管理丛书）

ISBN 978-7-111-56256-6

I. 移… II. 连… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字（2017）第 047684 号

本书根据作者近年来在移动数据挖掘方向的研究成果和工作进行编写，在主题上与当前学术界和工业界的热点相一致，自成体系，内容丰富，介绍了移动数据挖掘的基本概念和方法，包括移动数据预处理、用户移动模型、用户画像以及兴趣位置推荐等。

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：北京诚信伟业印刷有限公司	版　次：2017 年 5 月第 1 版第 1 次印刷
开　本：170mm×242mm　1/16	印　张：9.5
书　号：ISBN 978-7-111-56256-6	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

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

前 言

大约在十年前，本书作者所在的研究团队，也就是目前的微软亚洲研究院社会计算组，对挖掘人群移动数据中隐藏的知识产生了兴趣。这个团队在2007年开展了GeoLife项目，通过用户主动分享的移动数据来研究用户的出行模式，为旅游规划等应用提供支持。基于这个项目在WWW 2009大会上发表的论文“Mining interesting locations and travel sequences from GPS trajectories”目前引用数已经上千，在学术界产生了一定的影响。本书第一作者，目前在电子科技大学任教的连德富教授，长期针对基于移动数据的推荐系统进行研究，发表了大量有影响力的研究成果。

在过去十年，随着室内外定位、移动社交网络和物联网技术的发展与普及，移动数据的种类、规模和产生速度一直在迅速增加。这些数据中很大一部分是由人产生的，也就是通过各种方式记录下来的人的活动历史。它们包含了大量的知识，对于众多实际应用有着重要的价值。我们可以通过对这些数据进行挖掘，发现人类出行的规律，并针对用户的属性和兴趣爱好生成画像，从而为用户提供更加个性化的服务，包括交通出行规划、旅游线路和购物餐饮推荐等。这些知识还能用来研究疾病

传播、城市发展以及人类迁徙等具有重大社会意义的科学问题。

在实际应用中，移动数据的形式多种多样，既有来自移动社交网络的签到数据，来自运营商的日志数据，也有来自公交计费系统的刷卡记录数据，还有很多并不是由人产生的数据，例如由车辆、船舶甚至动物的移动生成的数据。在本书中，我们试图以人群移动数据为例，探讨和设计针对移动数据的数据挖掘算法，并指出在该领域展开研究将面临的挑战，希望这些经验也同样能应用到其它类型的移动数据上。

编辑为本书封面选取了清代苏州籍宫廷画家徐扬的巨幅长卷画作《姑苏繁华图》。在画中，画家通过自己对城市的理解，重现了苏州“商贾辐辏，百货骈阗”的市井风情。令人惊叹的是，据说全画中有各色人物 1 万 2 千余人。将他们的活动一一刻画出来是一个浩大的工程，反映了画家对苏州居民生活和出行规律的深刻理解，这也完美呼应了本书的主题。

最后，我们希望本书能帮助有兴趣研究移动数据挖掘的读者缩短学习的过程，共同推进该领域的研究进展。

作者

2017 年 4 月

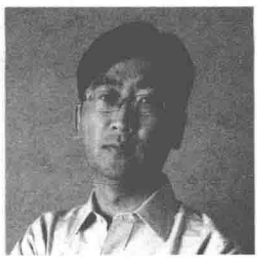
连德富 博士，电子科技大学计算机学院讲师，教育大数据研究所副所长。2009 年本科毕业于中国科学技术大学计算机学院，2014 年在中国科学技术大学获得计算机应用专业博士学位，获得中科院院长奖。他的研究领域包括用户建模、推荐系统、时空数据挖掘、教育大数据等等，在 KDD、IJCAI、WWW、ICDM、TIST 等顶级期刊和会议上发表 20 余篇论文，相关研究被麻省理工评论、中国青年报、中国科学报等媒体多次报道。他多次担任国际顶级会议的程序委员会委员和顶级期刊的审稿人，是 ACM、IEEE 和 CCF 会员。

张富峥 博士，于 2015 年 7 月加入微软亚洲研究院，现任社会计算组副研究员。他于 2015 年在中国科学技术大学获得计算机应用专业博士学位，于 2010 年在中国科学技术大学分别获得计算机和统计与金融学士学位。他是 ACM 和 IEEE 会员。他的研究领域包括了用户模型、推荐系统、深度学习、情感检测、社交网络、时空数据挖掘、普适计算、大规模系

统等方向。他在人工智能领域的重要会议和期刊上发表 20 余篇文章，如 KDD、WWW、Ubicomp、TIST 等，曾获 ICDM2013 最佳论文大奖，并多次担任人工智能领域大会的特邀审稿人和执行委员会委员。

王英子 中国科学技术大学和微软亚洲研究院联合培养博士生，于 2013 年获得中国科学技术大学计算机学士学位。她于 2014 年至 2016 年作为全职实习生在微软亚洲研究院社会计算组进行科学研究，研究领域包括时空数据挖掘、普适计算、推荐系统和时间序列预测等，在 KDD、Ubicomp、ICDM、KAIS 等会议和期刊上发表若干篇论文，曾经获得 ICDM 最佳论文奖和 KDD 最佳学生论文奖，并且多次担任国际顶级会议的审稿人。

袁晶 博士，现任微软高级科学家，Bing 中国应用科学团队负责人，此前为微软亚洲研究院研究员。他于 2007 年本科毕业自中国科学技术大学少年班学院，专业为计算数学，后于 2012 年获得该校计算机软件与理论博士学位，师从陈国良院士。他已在数据挖掘、普适计算、地理空间系统等领域的国际顶级会议和期刊上发表了 50 余篇论文，并多次获得包括 SIGSPATIAL、ICDM、KDD 在内的国际会议的最佳论文奖项。他曾作为程序委员会主席或领域主席参与组织了多次国际学术会议，并长期担任多个国际顶级会议(如 KDD、WWW、AAAI)的程序委员会委员。他于 2011 年被评为微软学者，现为 ACM 和 IEEE 会员。



谢幸 博士，于 2001 年 7 月加入微软亚洲研究院，现任社会计算组高级主任研究员，并任中国科技大学兼职博士生导师。他分别于 1996 年和 2001 年在中国科技大学获得计算机软件专业学士和博士学位。目前，他的团队在数据挖掘、普适计算和社会计算等领域展开创新性的研究。他在

国际会议和学术期刊上发表了200余篇学术论文，共被引用13000余次，H指数为54，并多次在KDD、ICDM等顶级会议上获最佳论文奖。他是ACM、IEEE高级会员和计算机学会杰出会员，多次担任顶级国际会议程序委员会委员和领域主席等职位。他是ACM Transactions on Intelligent Systems and Technology、Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)、Springer GeoInformatica、Elsevier Pervasive and Mobile Computing等杂志编委。他参与创立了ACM SIGSPATIAL中国分会，并曾担任ACM UbiComp 2011大会程序委员会共同主席。

目 录

第1章

引　言

移动数据挖掘研究的是基于移动数据的数据挖掘算法。这些数据挖掘算法需要更多地利用移动数据的特性，挖掘与这些特性有关系的模式。比如，研究发现，移动数据通常具有空间的聚集效应，即人们总是在少数的几个地点(家、工作场所等)附近活动，因而如何在数据挖掘的过程中考虑这一特性，便是移动数据挖掘需要重点考虑的一个问题。那么，移动数据具体是什么、有哪些特性、移动数据挖掘有什么任务、将要面对哪些挑战呢?

1.1　移动数据及其价值

移动数据是移动轨迹的集合，而移动轨迹可以简单地认为是移动记录的有序序列，既可以是人的移动数据，也可以是任何其他动物的移动数据。本书关注的是人类的移动历史。人类的移动历史具有更多的不确定性，他们并非总是愿意保持固定不变的生活规律，因而人类的移动数据中具有更加丰富的移动模式。人们可以通过携带 GPS 设备直接收集移动数据，也可以将诸如出租车、公交车、飞机、火车等移动对象作为载

体来间接收集他们的移动数据。这种移动数据收集的普适性得力于移动通信和传感设备等位置感知技术的发展和智能移动设备的普及，使得移动对象无论身处室内还是室外都可以更加容易地获取他们自身的地理位置信息。目前最先进的定位系统不仅依赖于全球卫星定位系统的高精度定位，还依赖于 Wi-Fi 和基站的较为粗略但范围更广的定位。出于业务本身或未来业务扩展及研究的需要，移动对象的很多定位数据都会被保留下来。由于与业务的强相关性，用户群的大小及位置的采样频率也决定了这些存留的位置数据不仅数量巨大，而且数据产生的速率很高。比如，运营商出于高效通信的需求会记录每个移动用户的服务位置，由于用户的规模巨大，因此每天产生的位置数据量也是非常巨大的。据我国三大运营商的运营数据显示，截至 2015 年 12 月，中国电信、中国移动和中国联通的用户数分别高达 1.979 亿户、8.26 亿户和 2.866 亿户。假如每人每天平均通信一次，那么每天就会有约 13 亿条的位置数据。

然而，正如基站定位数据是存储在运营商手中的一样，位置数据一般不会保存在移动对象的手中，外加数据量巨大，使得移动数据的开放受到了很大的约束。不过，随着移动互联网和在线社交网络的发展，诸如街旁网、Foursquare、Facebook Place 等位置社交网络应运而生。在位置社交网络中，人们可以便捷地跟踪和分享诸如他们在什么地方和什么时候做了什么事情的签到(check-in)记录等位置访问信息。同样，源于与在线社交网络的结合，位置社交网络中的用户群也是巨大的，使得用户的移动数据也得到了大量的积累。根据街旁网的官方数据，从 2010 年 5 月上线到 2013 年 7 月，街旁用户数已经突破了 500 万，累计签到次数超过 8000 万次。根据 Foursquare 的统计数据，从 2009 年 3 月上线到 2013 年 12 月，用户数已经达到了 4500 万，累计签到数高达 50 亿。

这些大规模移动数据的积累，为基于位置的智能服务提供了重要的基础条件。目前，这些基于位置的智能服务开辟了一个正在快速增长的市场。一份来自 MarketsandMarkets 的研究报告预计，诸如导航、移动广告、移动社交网络等基于位置的智能服务的市场份额将从 2016 年的