

掘金大数据

电信数据金矿详解、挖掘及应用

主编 / 程新洲 朱常波 晁昆

副主编 / 李海彬 贾玉玮

崔曙光
深圳市大数据研究院副院长、
中国工程院院士

唐雄燕
中国联通网络技术研究院首席科学家

刘红杰
博识创智执行董事、总经理

联袂推荐

张涌
中国联通网络技术研究院院长

倾心指导

刘韵洁
中国工程院院士

作序推荐



机械工业出版社
CHINA MACHINE PRESS

掘金大数据：电信数据 金矿详解、挖掘及应用

主 编 程新洲 朱常波 晁昆
副主编 李海彬 贾玉玮

机械工业出版社

在数据横向融合的时代，充分挖掘数据金矿及盘活数据资产，是企业发展和转型的关键所在。电信运营商以其数据特殊性，必将成为大数据领域的领航者、生力军。各行业的大数据从业者要如何从电信业的大数据中挖掘价值呢？

本书彻底揭开电信运营商数据的神秘面纱，系统介绍了大数据的发展历程，主要的数据挖掘方法，电信运营商在网络运行及业务运营方面的数据资源特征，基于用户、业务、网络、终端及内在联系的电信运营商大数据分析体系，以及该分析体系在企业内部及外部行业的应用实践。本书特别列举了一系列电信大数据在企业运营以及多个行业应用的案例，为各行业的大数据应用推进提供参考，同时对于大数据推进过程中的各种风险与挑战提出应对措施。

对于从事电信运营商大数据分析工作的读者，本书是一本难得的实务指南；对于从事电信运营商运营工作的管理人员，本书提供了运营和决策的辅助参考手段；对于其他行业的大数据从业人员，本书以电信运营商大数据应用为导向，介绍的案例涵盖多个领域（交通、金融、旅游、广告、气象、餐饮、娱乐、医疗、政务、城市规划等），更有相关产品介绍，涉及的应用思路、方法、解决方案能给予读者很多启发和思考。

图书在版编目（CIP）数据

掘金大数据：电信数据金矿详解、挖掘及应用/程新洲，
朱常波，晁昆主编。—北京：机械工业出版社，2018.10
ISBN 978-7-111-61113-4

I. ①掘… II. ①程… ②朱… ③晁… III. ①电信运营企业-
数据管理-研究 IV. ①F626.1

中国版本图书馆 CIP 数据核字（2018）第 228322 号

机械工业出版社（北京市百万庄大街 22 号 邮政编码 100037）

策划编辑：刘 洁 责任编辑：刘 洁

责任校对：舒 莹 责任印制：常天培

北京圣夫亚美印刷有限公司印刷

2019 年 1 月第 1 版·第 1 次印刷

170mm×242mm·16.5 印张·1 插页·249 千字

标准书号：ISBN 978-7-111-61113-4

定价：59.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务

服务咨询热线：010-88361066

读者购书热线：010-68326294

010-88379203

封面无防伪标均为盗版

网络服务

机工官网：www.cmpbook.com

机工官博：weibo.com/cmp1952

金书网：www.golden-book.com

教育服务网：www.cmpedu.com

编 委 会

Editorial Committee

编委会主席：张 涌

主 编：程新洲、朱常波、晁昆

副 主 编：李海彬、贾玉玮

委 员：（排名不分先后）

袁明强、张 恒、李明欣、刘媛妮、

宋春涛、高 洁、关 键、徐乐西、

曹晓冬、张 涛、曹丽娟、王永峰、

慕明君、任鹏飞、苏 飞、杨 杉、

张 帆、魏东晓、成 晨、王鹏飞、

叶海纳、韩玉辉、董润莎、徐争莉

序

Preface

随着科技的进步和发展，电脑、手机、车载设备等生产或传播数据的载体不断增多，带来全球数据的急速增长，据 IDC（国际数据公司）预计，到 2020 年全球的数据总量将达到 40ZB。如今，数据已经渗透到当今每一个行业和业务领域，成为一种可再生的新资源。数据也已成为各个行业的重要生产力，助力各领域通过数据洞见本质、预测发展趋势、指引未来决策、实现资源优化配置。

2015 年，十八届五中全会首次提出“国家大数据战略”，《促进大数据发展行动纲要》发布。2017 年 12 月 8 日，习近平总书记在中共中央政治局第二次集体学习时强调，大数据发展日新月异，我们应该审时度势、精心谋划、超前布局、力争主动，推动实施国家大数据战略，加快完善数字基础设施，推进数据资源整合和开放共享，保障数据安全，加快建设数字中国，更好服务我国经济社会发展和人民生活改善。

电信运营商拥有遍布全国的通信网络，为人与人之间的通信提供语音和数据服务，同时也客观、真实地记录和存储着每一条信令和用户行为信息，数据体量巨大，内容也十分丰富。近几年来，中国联通着力全网大数据的标准化采集、解析、平台建设及挖掘应用工作，在企业内部的网络建设、市场营销、客户维系等多个方面实现了降本增效和精细化运营；同时致力于相关行业的跨界合作，推动数据资源融合和开放共享，促进大数据产业生态的不断完善，构建全方位社会服务的智能化。

中国工程院院士



前 言

Foreword

近年来，随着科技的发展和进步，各行各业信息化程度逐渐提高，由此产生的海量数据正处于爆炸式增长态势。越来越多的人意识到，海量数据中蕴含着丰富的价值信息，如何融合多种数据，充分挖掘数据金矿及盘活数据资产，是企业发展和转型的关键所在。

电信运营商作为各类新兴业务及运营模式的载体，“管道化”趋势日益显现，并从中积累了丰富的数据资源。电信运营商数据具有覆盖范围广、数据准确度高、数据持续性长等优势，其金矿价值不断凸显，由服务运营转为数据运营已经成为电信运营商的必然选择。本书由中国联通网络技术研究院高级专家程新洲及其团队创作完成，所有内容均来自编者长期工作经验的总结与凝练，旨在为从事运营商大数据相关工作的人员以及对运营商大数据感兴趣的各行各业人员提供数据分析的思路和方法。

本书第1章概括性地介绍了大数据的发展历程和时代背景。之后，本书将重点落在电信运营商大数据上，在第2章全面介绍了运营商BSS（业务支撑系统）域和OSS（运营支撑系统）域共21类典型数据源的主要内容、特点及来源。第3章系统描述了运营商大数据分析体系的构建过程，以用户画像、用户轨迹、用户偏好、业务特征捕获、用户业务感知评价、网络全景分析、终端分类构成解析等为链条，从用户、业务、网络、终端、运行匹配度、众筹竞争力角度实现了多维度洞察以及全方位综合评判。第4章和第5章分别介绍了大数据在电信运营商企业内部及外部的应用方向。在企业内部将用户业务诉求和网络服务能力进行了匹配，使大数据分析作为新时代背景下网络规划建设、运维优化、市场推送以及客户维系的新手段。在

企业外部，结合应用实践，详细阐述了大数据在诸如交通、金融、城市规划、旅游、广告、公众及气象、餐饮、娱乐、政务等行业的应用案例及产品解决方案。最后，在第6章中对大数据面临的风险与挑战进行了总结及展望。

对于从事运营商大数据分析工作的人来说，本书是一本难得的实务指南；对于其他行业大数据从业人员来讲，本书贴近运营商现状，深度、广度适中，是一本非常好的阅读材料。同时，书中涉及的应用思路和方法相信能够给予读者很多的启发和思考。

鉴于编者水平有限，时间较紧，且大数据技术发展及应用日新月异，本书在编写过程中难免存在疏漏及不足之处，恳请读者批评指正。

编者

2018年9月

目 录

Contents

编委会
序
前言

第1章 大数据的发展历程和时代背景 //1

- 1.1 大数据发展的五大驱动力 //2
- 1.2 大数据发展的历程及技术演进 //4
 - 1.2.1 采集解析技术 //4
 - 1.2.2 存储管理技术 //6
 - 1.2.3 并行计算技术 //8
- 1.3 大数据引起的时代变革 //9
 - 1.3.1 智能派单、路线优化 //9
 - 1.3.2 金融授信、风险防范 //10
 - 1.3.3 广告精准投放 //10
 - 1.3.4 挖掘用户的隐性需求 //10
- 1.4 电信大数据的掘金之路 //11
 - 1.4.1 AT&T //12
 - 1.4.2 德国电信 //12
 - 1.4.3 Vodafone //12
 - 1.4.4 法国电信 //12
 - 1.4.5 中国联通 //12

第2章 电信运营商的价值数据 //14

- 2.1 电信运营商数据源概览 //14

- 2.2 网络运行数据 //17
 - 2.2.1 基础资源及配置数据 //17
 - 2.2.2 信令追踪数据 //21
 - 2.2.3 业务识别数据 //25
 - 2.2.4 性能统计数据 //34
 - 2.2.5 监控预警数据 //39
- 2.3 业务运营数据 //43
 - 2.3.1 用户基础资料 //43
 - 2.3.2 用户业务行为 //44
 - 2.3.3 用户辅助信息 //46

第3章 电信大数据分析体系 //50

- 3.1 用户画像及行为洞察 //53
 - 3.1.1 用户全息画像构建 //53
 - 3.1.2 用户实时轨迹追踪 //61
 - 3.1.3 用户行为偏好分析 //70
- 3.2 业务识别及感知评判 //80
 - 3.2.1 业务特征捕获 //80
 - 3.2.2 用户业务感知评价 //85
- 3.3 网络分析及全景描绘 //93
 - 3.3.1 网络资源分析 //93
 - 3.3.2 网络覆盖分析 //96
 - 3.3.3 网络性能分析 //100
 - 3.3.4 网络结构分析 //104
- 3.4 终端解构及性能评价 //105
 - 3.4.1 终端分类构成解析 //105
 - 3.4.2 终端综合性能评价 //112
- 3.5 运行匹配度解析 //116
 - 3.5.1 终端与网络匹配度 //117

- 3.5.2 网络资源与业务匹配度 //118
- 3.6 众筹竞争力分析 //120
 - 3.6.1 传统分析方法及局限性 //120
 - 3.6.2 基于智能终端 App 数据的分析方法 //120
 - 3.6.3 分析结果示例 //122
- 第4章 电信大数据企业内部应用 //125**
 - 4.1 大数据助力资源精准投放——网络规划建设 //125
 - 4.1.1 网络规划资源的精准投放 //126
 - 4.1.2 网络建设效果的有效评价 //127
 - 4.2 大数据助力效率提升——网络运维优化 //129
 - 4.2.1 智能化运维 //129
 - 4.2.2 精细化优化 //133
 - 4.3 特定用户群体的精准营销——市场推送 //136
 - 4.3.1 流量经营时代 //137
 - 4.3.2 现网中价值区域的挖掘 //138
 - 4.3.3 基于 B+O 大数据的流量经营策略 //145
 - 4.3.4 新型业务营销及发展趋势 //156
 - 4.4 用户黏性及离网预判——客户维系 //161
 - 4.4.1 存量客户 //163
 - 4.4.2 现有用户的套餐升级 //168
 - 4.4.3 潜在异网客户 //169
- 第5章 电信大数据外部行业应用 //171**
 - 5.1 电信大数据在交通领域的应用 //171
 - 5.1.1 交通行业的发展需求 //172
 - 5.1.2 交通行业数据现状 //175
 - 5.1.3 电信运营商在交通领域的角色定位 //177
 - 5.1.4 电信大数据助力交通领域的应用案例 //181
 - 5.2 电信大数据在金融领域的应用 //183

- 5.2.1 金融行业发展需求 //183
- 5.2.2 金融行业数据现状 //184
- 5.2.3 电信运营商在金融行业的角色定位 //186
- 5.2.4 电信大数据助力金融的应用案例 //188
- 5.3 电信大数据在城市规划领域的应用 //193
 - 5.3.1 城市规划发展需求 //193
 - 5.3.2 城市规划数据现状 //193
 - 5.3.3 电信运营商在城市规划的角色定位 //194
 - 5.3.4 电信大数据助力城市规划的应用案例 //195
- 5.4 电信大数据在旅游领域的应用 //198
 - 5.4.1 旅游行业发展需求 //198
 - 5.4.2 旅游行业数据现状 //199
 - 5.4.3 电信运营商在旅游行业的角色定位 //199
 - 5.4.4 电信大数据助力旅游行业的应用案例 //201
- 5.5 电信大数据在广告领域的应用 //204
 - 5.5.1 广告行业发展需求 //204
 - 5.5.2 广告行业数据现状 //204
 - 5.5.3 电信运营商在广告行业的角色定位 //206
 - 5.5.4 电信大数据助力广告精准投放的应用案例 //207
- 5.6 电信大数据在公众及气象领域的应用 //210
 - 5.6.1 公众及气象领域发展需求 //211
 - 5.6.2 公众及气象领域数据现状 //212
 - 5.6.3 电信运营商在公众及气象领域的角色定位 //213
 - 5.6.4 电信大数据助力公众及气象领域的应用案例 //214
- 5.7 电信大数据在餐饮、娱乐领域的应用 //216
 - 5.7.1 餐饮、娱乐行业发展需求 //217
 - 5.7.2 餐饮、娱乐行业数据现状 //218
 - 5.7.3 电信运营商在餐饮、娱乐行业的角色定位 //219

- 5.7.4 电信大数据助力餐饮、娱乐行业的应用案例 //220
- 5.8 电信大数据在其他领域的应用 //223
 - 5.8.1 商业决策 //223
 - 5.8.2 治安监控 //225
 - 5.8.3 政府决策 //226
 - 5.8.4 医疗行业 //228
 - 5.8.5 教育行业 //230
- 第6章 大数据的风险与挑战 //232**
 - 6.1 风险管理 //232
 - 6.1.1 数据泡沫风险 //232
 - 6.1.2 技术升级风险 //235
 - 6.1.3 隐私侵权风险 //240
 - 6.2 挑战应对 //244
 - 6.2.1 技术发展挑战 //244
 - 6.2.2 思维转变挑战 //246
 - 6.2.3 商业挑战 //248
- 参考文献 //251**

第 1 章 大数据的发展历程和时代背景

“大数据”又称巨量资料，指的是所涉及的数据资料量规模巨大到无法通过人脑甚至主流软件工具，在合理时间内达到撷取、管理、处理并整理成为对企业经营决策起到更积极作用的资讯。

从技术上看，大数据与云计算的关系就像一枚硬币的正反面一样密不可分。大数据必然无法用单台计算机进行处理，而必须采用分布式计算架构。大数据技术依托云计算的分布式处理、分布式数据库、云存储和/或虚拟化技术，从而进行海量数据的分析及挖掘。在维克托·迈尔-舍恩伯格（Viktor Mayer-Schönberger）及肯尼斯·库克耶（Kenneth Cukier）编写的《大数据时代：生活、工作与思维的大变革》一书中，介绍了大数据的 4V 特点：Volume（大量）、Velocity（高速）、Variety（多样）和 Value（价值）。

（1）数据体量巨大（Volume）。截至目前，人类生产的所有印刷材料的数据量是 200PB，而历史上全人类说过的所有话的数据量大约是 5EB。当前，典型个人计算机硬盘的容量为 TB 量级，而一些大企业的数据量已经接近 EB 量级。字节的次方单位如表 1-1 所示。

表 1-1 字节的次方单位

中文名字	英文名字	缩写	次方
千字节	Kibibyte	KB	10^3
兆字节	Mebibyte	MB	10^6
吉字节	Gibibyte	GB	10^9
太字节	Tebibyte	TB	10^{12}
拍字节	Pebibyte	PB	10^{15}
艾字节	Exbibyte	EB	10^{18}
泽字节	Zebibyte	ZB	10^{21}
尧字节	Yobibyte	YB	10^{24}

(2) 数据类型繁多 (Variety)。这种类型的多样性也让数据被分为结构化数据和非结构化数据。相对于以往便于存储的以文本为主的结构化数据，非结构化数据越来越多，包括网络日志、音频、视频、图片、地理位置信息等，这些多类型的数据对数据的处理能力提出了更高要求。

(3) 价值密度低 (Value)。价值密度的高低与数据总量的大小成反比。以视频为例，一部 1 小时的视频，在连续不间断的监控中，有用的数据可能仅有一二秒。如何通过强大的机器算法更迅速地完成数据的价值“提纯”成为目前大数据背景下亟待解决的难题。

(4) 处理速度快 (Velocity)。这是大数据区别于传统数据挖掘的最显著的特征。根据 IDC (国际数据中心) 的“数字宇宙”报告，预计到 2020 年，全球数据使用量将达到 40ZB。在如此海量的数据面前，处理数据的效率就是企业的生命。

近几年来，随着计算机和信息技术的迅猛发展，各行各业应用系统的规模迅速扩大，行业应用所产生的数据呈爆炸性增长。动辄数百 TB 甚至数百 PB 的大数据已远远超出了现有传统的计算技术和信息系统的处理能力，因此，寻求有效的大数据处理技术、方法和手段已经成为现实世界的迫切需求。

1.1 大数据发展的五大驱动力

近几年，互联网行业发展风起云涌，“大数据”炙手可热，大数据的发展是一个过程，大数据的应用才是最终目的。回顾大数据产业发展的驱动力，主要有如下 5 种：

第一，是政府产业引导的驱动力。政府由于行政管理的原因，积累了大量的数据。《促进大数据发展行动纲要》中明确指出，到 2018 年年底，建成国家政府数据统一开放平台。政府数据的开放共享是推动大数据产业发展最重要的驱动力。我们常用的百度地图等丰富多彩的服务就依赖于国家地图数据的开放。

第二，是金融资本的驱动力。大数据产业必然要求数据量大、用户量大，在数据与用户未能达到引爆市场之前，需要大体量资金的支持。因此，金融资本是大数

据产业快速发展的驱动力。如某网约车公司，在2012年12月，刚成立半年就获得了300万美元的融资，2013年4月完成B轮融资，获得腾讯1500万美元的投资。大体量资金的介入使其通过补贴司机与用户快速获得网约车App 59.4%市场份额的原因。2014年成立的贵州省食品安全云在其天使轮获得2300万元人民币的投资，迅速形成了20余款面向政府、食品企业、检验检测机构和消费者的产品和服务，初步形成了服务于食品安全社会共治、几类用户协同发展的平台架构。

第三，是技术发展的驱动力。一些掌握先进信息技术的公司在发展中，快速积累了用户和数据，推动了大数据产业的发展。例如，2000年1月创立于北京中关村的百度，掌握着先进的搜索引擎技术，面对用户的搜索产品不断丰富，迅速成为全球最大的中文搜索引擎；成立于1998年11月的腾讯，掌握着即时通信技术，在三年的时间里，公司的产品QQ的注册用户数突破1亿人，形成了用户大数据，如今腾讯还推出了腾讯云分析（Mobile Tencent Analytics，MTA）、腾讯云推送（信鸽）等大数据运营服务。

第四，是产品应用的驱动力。2014年，阿里巴巴对外宣称已经拥有100PB数据并以令人欣喜的速度增长，阿里巴巴战略定位为云端+大数据。凭借阿里系产品的众多应用，阿里巴巴大数据的来源覆盖人们生活的各个层面。最初通过淘宝电商平台，阿里巴巴采集了商户和用户基于买和卖的商品信息、商家信息、消费者信息（如姓名、地址、电话、消费习惯、购买能力等数据），后来随着支付宝的应用，阿里巴巴采集了用户银行信息、身份证信息等，形成了对用户全面形象刻画的数据。可以说，正是产品大范围的推广应用，电子商务的快速发展，促进了阿里大数据产业的发展。

第五，是资源集聚的驱动力。除了类似阿里巴巴这类用户与商品等资源集聚的大数据产业外，一些禀赋独特的资源集聚也能形成大数据的驱动力。例如交通，因为车辆的集聚而形成大数据；旅游，因为旅游资源的集聚而形成大数据；科技，因为科技要素的集聚而形成大数据；人力，因为人才信息的集聚而形成大数据……挖掘各类集聚的数据资源，才能形成大数据产业快速发展的驱动力。

大数据涉及社会生活的方方面面，任何一个方面密集的信息需求都可能形成大

数据产业发展的驱动力。除了上述 5 大驱动力外，可能还有很多驱动力。找到发力点，也就探明了大数据产业的商业模式，将推动大数据产业快速发展。

1.2 大数据发展的历程及技术演进

大数据技术是指从各种各样类型的巨量数据中快速获得有价值信息的技术。解决大数据问题的核心是大数据技术。目前所说的“大数据”不仅指数据本身的规模，也包括采集数据的工具、平台和数据分析系统。大数据研发的目的是发展大数据技术并将其应用到相关领域，通过解决巨量数据处理问题促进大数据突破性发展。因此，大数据时代带来的挑战不仅体现在如何处理巨量数据而从中获取有价值的信息，也体现在如何加强大数据技术研发，抢占时代发展的先机。

1.2.1 采集解析技术

大数据具有“体量大、结构多样、时效性强”等特征，要用好各类大数据，对数据源进行数据准备处理是必备的前置环节。电信运营商大数据面对着复杂的生产环境：设备厂商多、网元种类繁多、数据类型及字段复杂、字段定义与电信技术交缠。传统数据处理体系中的 ETL（抽取、转换、加载）前置数据处理已不足以完成电信运营商复杂环境中对数据的解析、关联、标准化及准确性要求。因此，数据采集解析关键技术研究方面，需充分研究网元设备原始数据提供能力、数据接口方式、原始数据解析及标准化、跨行业数据融合、多层次数据关联等数据采集解析技术，在传统 ETL 前置数据处理的基础上，根据电信运营商大数据面广量大、动态复杂的特点，将网络原始数据梳理解析，初步加工为可用的规整且意义明确的元数据。这方面的工作是大数据技术构架的基础。此外，积极寻找并发掘网络数据外的其他关联数据源，并经采集解析相关的前置处理后引入大数据系统中，也是采集解析技术研究需要关注的研究方向。

1. 关注网元设备原始数据提供能力

通信技术在不断演进发展，通信设备也随之进行代际更替，数据应用对原始数据的粒度、范围、类型等都不断提出更高的要求。例如使用 NFV（Network Function

Virtualization, 网络功能虚拟化) 技术后新出现的虚拟网元、物联网设备、5G (The 5th Generation, 第五代移动通信技术) 新设备等; 通信标准升级也带来了新通信功能技术点, 如 VoLTE (Voice over LTE, 基于 LTE 的语音) 技术、最小化路测技术、SON (Self - Organizing Network, 自组织网络) 技术等; 上层数据应用也对原始数据提出新需求, 如对 MR (Measurement Report, 测量报告) 的定位需求、对数据采集周期缩小的需求等。电信运营商需深入研究并向设备厂商提出原始数据采集需求, 推动设备厂商进行原始数据开发, 及时准确地从设备上提供原始数据。

2. 研究数据接口技术

随着 IT 技术的发展, 数据接口技术也在不断演进, 如 CORBA (Common Object Request Broker Architecture, 公共对象请求代理体系结构) 接口的简化、FTP (File Transfer Protocol, 文件传输协议) 的高效加密、流式消息处理等技术出现后, 为数据采集接口升级改造提供了技术条件。电信运营商需具体研究各类数据需求, 研究从厂商网元设备、OMC (Operation and Maintenance Center, 操作维护中心)、链路接口等多途径获取应用所需数据的接口方式, 提出对现有不满足应用需求的接口进行改造的技术方案。

3. 原始数据解析及标准化研究

根据数据特点及上层应用的需求, 对多厂商、多种类的异源数据进行解析及标准化, 使其规整统一, 以利于上层应用。在解析及标准化的过程中, 如何在尽量保持原有语义的情况下去粗取精、消除噪声, 形成规范化的可以标准方式利用的数据词典, 将是研究的重点。

4. 跨行业数据源融合研究

通过项目合作、引用公开数据等方式了解电信运营商范围外的其他大数据源或大数据产品, 如交通路网流量、搜索点击等统计数据、地图及 POI (Point of Interesting, 兴趣点) 数据、天气气候数据等, 研究探索使用外部大数据资源为运营商网络规划、建设、维护、优化等提供应用工具的技术。