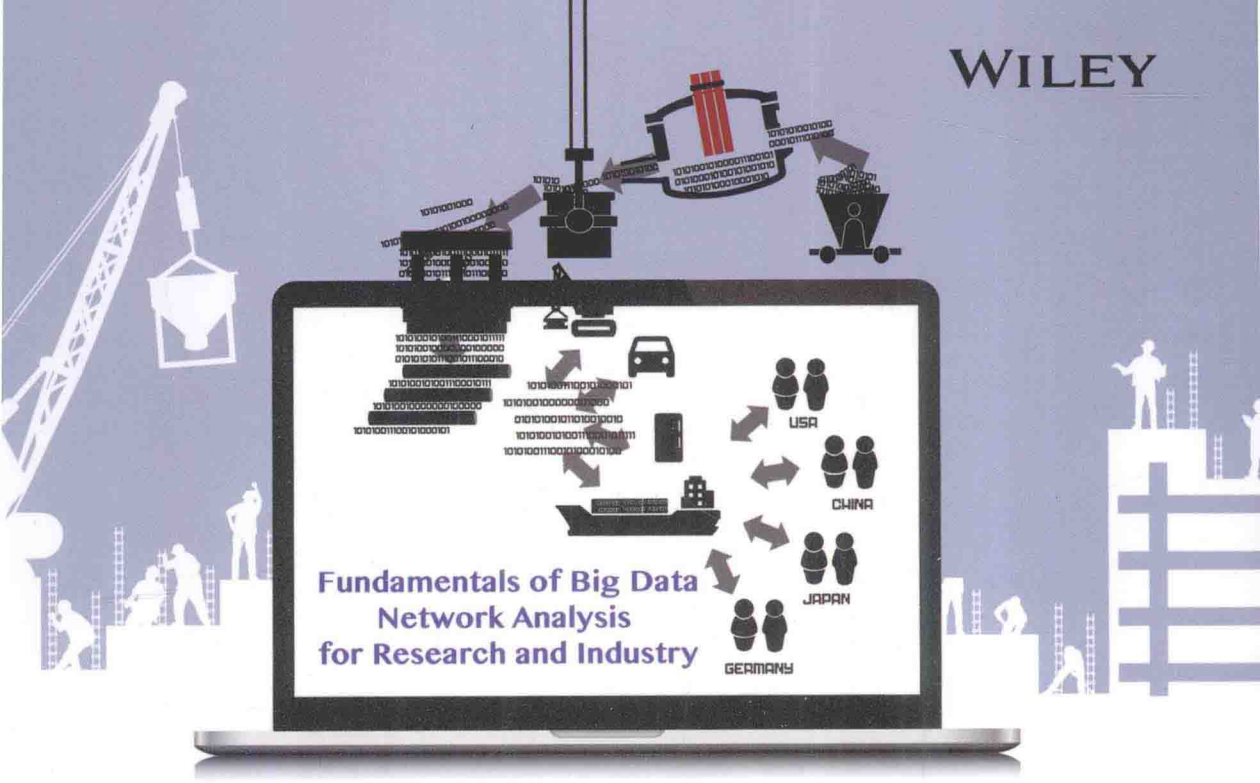
WILEY



大数据科学与应用丛书

工业大数据实践
工业4.0时代
大数据分析技术与实践案例

[新加坡] Hyunjoung Lee（李贤荣） 著
Il Sohn（孙 宁）

向 阳 刘让龙 寇晶琪 译

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

大数据科学与应用丛书

工业大数据实践

工业 4.0 时代大数据分析技术与实践案例

Fundamentals of Big Data Network Analysis for Research and Industry

[新加坡] Hyunjoung Lee（李贤荣）
Il Sohn（孙宁） 著

向 阳 刘让龙 寇晶琪 译

電子工業出版社
Publishing House of Electronics Industry
北京 • BEIJING

内 容 简 介

如今，海量的数据无处不在，从数据中提取关键信息的能力显得愈发重要。本书从崭新的视角认识大数据，研究了钢铁行业中的典型大数据案例，为读者提供进行数据网络分析、数据中有效信息提取的详细步骤和指导方法。特别是在网络分析方法方面，对数据采集、研究方法设计及分析、数据结果呈现进行了介绍。同时，介绍了相关网络分析软件：UCINET、NetMiner、R、NodeXL 及 Gephi。

本书适合分析师、研究工程师、工业工程师、市场营销专家，以及对大数据分析感兴趣的人员阅读与参考。

图书在版编目（CIP）数据

工业大数据实践：工业 4.0 时代大数据分析技术与实践案例/（新加坡）李贤荣，（新加坡）孙宁著；向阳，刘让龙，寇晶琪译. —北京：电子工业出版社，2017.6

（大数据科学与应用丛书）

书名原文：Fundamentals of Big Data Network Analysis for Research and Industry

ISBN 978-7-121-31575-6

Ⅰ. ①工… Ⅱ. ①李… ②孙… ③向… ④刘… ⑤寇… Ⅲ. ①数据处理－应用－制造工业－研究 Ⅳ. ①F407.4-39

中国版本图书馆 CIP 数据核字（2017）第 108189 号

策划编辑：李树林

责任编辑：谭丽莎

印　　刷：三河市鑫金马印装有限公司

装　　订：三河市鑫金马印装有限公司

出版发行：电子工业出版社

北京市海淀区万寿路 173 信箱　邮编　100036

开　　本：720×1000　1/16　印张：13.75　字数：203 千字

版　　次：2017 年 6 月第 1 版

印　　次：2017 年 6 月第 1 次印刷

印　　数：3 000 册　　定价：49.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：（010）88254463；lisl@phei.com.cn。

译者序

当接到翻译的工作，看到详细的代码和图示时，我不禁又回忆起读博士期间一行行敲代码的生活。毕业后我虽不直接从事数据编程工作，但也一直进行着大数据产业发展的相关研究。作为一本技术层面的实操性书籍，本书的确写得非常详细，从概念到软件操作、从数据分析方法到实际案例剖析，一步步帮助读者掌握大数据理解和分析。书中花了大量篇幅为读者介绍主流大数据分析软件的操作和应用，很适合作为一本入门级的大数据工具书籍。

作为长期研究中国大数据产业趋势的分析师，我对于本书中提到的网络分析法感触颇深。目前很多企业在做大数据，大致都是从标准化的数据采集分析系统起步，大量的工作仍然集中在企业内部数据的整合上，而对于外部消费者数据和内部企业流程数据的连通融合，则是未来亟待解决的关键性问题。本书从全新的数据网络关系视角入手，为我们清晰展现了从数据采集、数据清理、数据分析到数据可视化的全流程步骤。记得在最新一季的《黑镜》中，有一集就是未来社交网络数据的智能化，通过人群在社交网络上的言论统计来操作机器人。这种科幻剧中的场景恰好与本书中分析 Facebook、Twitter 上的网络关联数据不谋而合。可以说，社交数据正在成为大数据分析中不可或缺的一环。

同时，在国内产能过剩的大环境下，有关钢铁、煤炭的大数据应用也是未来的焦点之一。如何通过大数据来提升产品质量、发掘更精准的市场需求成为钢铁煤炭企业的转型重点。目前国内有关这方面的大数据书籍较少，本

书从国际钢铁贸易的案例出发，为读者和行业专家深入剖析了大数据在工业领域的应用效果，并结合了不同的主流分析软件的详细使用教程，必将是各行业研究人员的得力助手。

向阳

前言

PREFACE

本书的理念最初是由未来钢铁技术论坛发起并支持的。在这个论坛上，一批未来钢铁技术的研究者们齐聚一堂，提出要在全球钢铁贸易区之间挖掘钢铁技术及产品植入的战略意义。在韩国钢铁及钢铁协会的赞助下，作者首次针对钢铁贸易数据进行分析，涵盖了贸易国之间的网络关系及跨境交易的钢铁产品信息。从最开始，该书作者就致力于通过钢铁贸易市场的一些案例向社会公众、行业研究员及数据分析专业的学生提供大数据分析的方法论。

本书共分为 8 章。第 1 章主要定义了什么是大数据及在企业内部管理中如何运用它来激发更多的产能和更高的效率。第 2 章介绍了大数据分析相关的各种不同软件，可以帮助识别目前市场上在售的分析软件的优缺点。第 3 章主要围绕社会网络分析进行介绍，给出了数据间网络关系结构中的节点和链接的定义。第 4 章总结了网络分析的研究方法论，包括设定一项实验、数据如何采集及如何过滤无效或干扰数据。第 5 章着重描述了中心性分析和凝聚子群分析，其中中心性分析包括中心度指标、中介中心性及亲近中心性。第 6 章对全书进行了总结，提出了网络的性能及节点对（或者数据对）之间的对等性，还重点概述了节点之间的连通性。第 7 章对 NetMiner 的数据结构进行了介绍。第 8 章对 NetMiner 中提供的样例数据进行网络分析。

经过 8 章的详细介绍，我们已经能够充分理解正在进行的大数据分析。书中提到的各种不同的分析方法和程序都是目前使用率最高的。本书旨在为初次接触大数据的学者或有部分基础的学者，全面介绍大数据涉及的基

础知识，以及上述学者在将来从事大数据实验时可能用到的分析方法。作者的众多朋友也为本书的顺利完成贡献了不小的力量。

在此，我们要向 Dong Joon Min(董炯民)教授表达真挚的感激，感谢他在钢铁数据分析中极具帮助性的独到见解。同时，感谢 Jae Wook Ryu(在旭柳)博士，感谢他一直以来提供的帮助；感谢 Doo-Hee Lee(杜河力)教授的鼓励及对学术的执着追求。

谨以本书献给我们的家人，感激他们在本书的写作过程中做出的牺牲和支持。

作　者

目录

CONTENTS

第 1 章　大数据从何而来　/ 1

1.1　大数据　/ 2
1.2　是什么产生了大数据　/ 7
1.3　我们如何利用大数据　/ 10
1.4　大数据相关的几个重要问题　/ 15
参考文献　/ 17

第 2 章　网络关系数据分析的基础工具　/ 19

2.1　UCINET　/ 20
2.2　NetMiner　/ 25
2.3　R　/ 31
2.4　Gephi　/ 35
2.5　NodeXL　/ 40
参考文献　/ 42

第 3 章　了解网络分析　/ 43

3.1　定义社会网络分析　/ 44
3.2　SNA 的基本概念　/ 46
3.3　社交网络数据　/ 50
参考文献　/ 54

第 4 章　采用 SNA 的研究方法　/ 55

4.1　SNA 实验程序　/ 56
4.2　识别实验问题和建立假设　/ 58
4.3　研究设计　/ 61
4.4　网络数据的获得　/ 67
4.5　数据清理　/ 73
参考文献　/ 83

第 5 章　位置和结构　/ 84

5.1　位置　/ 85
5.2　凝聚子群体　/ 103
参考文献　/ 109

第 6 章　连通性和角色　/ 111

6.1　连接分析　/ 112
6.2　角色　/ 119
参考文献　/ 129

第 7 章　NetMiner 的数据结构　/ 131

7.1　数据示例　/ 132
7.2　主要概念　/ 135
7.3　数据处理　/ 144
参考文献　/ 153

第 8 章　使用 NetMiner 的网络分析　/ 154

8.1　中心地位和凝聚力子群　/ 155

8.2　连通性和等同性　/ 167
8.3　可视化和探索性分析　/ 176

附录 A　可视化　/ 184

A.1　弹性算法　/ 185
A.2　多维比例算法　/ 187
A.3　聚类算法　/ 188
A.4　分层算法　/ 189
A.5　圆弧算法　/ 190
A.6　简单算法　/ 191
参考文献　/ 192

附录 B　案例研究：钢铁研究的知识结构　/ 194

参考文献　/ 207

第1章

大数据从何而来

这个社会每天都会产生大量的数据。这些原始数据并未经过过滤，伴随着这些数据的迅速累积，也产生了大量的无用的干扰数据，这些干扰数据必须被剔除，从而确保接下来有效客观的实验分析。这就要求分析人员具备足够的能力，从这些大量原始数据中提取出正确的、有用的信息。通过这样从沙子里淘出真正的金子的方式，大数据分析可以帮助企业从一个相对狭小的切入口出发，最终获取一个更广阔的商业视角。因为大数据的重要性是基于思维的扩张，这无关乎数据的量是不是够大、积累速度是不是够快或者数据的种类是不是足够多样，而是我们将采用不同的视角和观念去分析这些数据。就好比，如果你想看到一片森林，你就需要爬到高高的山顶上，而不是离开这片森林。同理，如果想对大数据有更深入的了解，我们应该尝试上升到足够的高度，拓宽我们的视角，上升得越高，我们的视野才足够大。为从森林内部捕捉到森林的外貌，就应当采取一个不同的视角，这正是大数据产生的意义。

1.1 大数据

近些年，大数据领域已经引起了众人足够多的兴趣。高德纳（Gartner）作为全球顶级市场分析研究所之一，早已在 2012 年和 2013 年连续两年将大数据列为前十大战略技术工具[1]。并且，在 2014 年还将大数据和可操作分析工具（Actionable Analytics）并列作为公司智能化治理的重要战略技术[2]。不仅如此，每年一次的达沃斯世界经济论坛，全球首脑和经济大臣们都会会聚在此，共同探讨全球问题，其中，2012 年的达沃斯论坛再一次向世人强调了大数据的重要性，将其确定为对全球未来发展至关重要的十大科学技术之一[3]。全球目前正面临着经济危机，部分地区刚开始好转，除此之外还有气候变化、能源短缺、贫穷及局部地区安全等各种严峻的挑战，大数据的此次入选意味着解决这些全球问题的途径就是需要大量广泛的数据，并且还急需通过有效的管理和提取有效的数据，来协助深入了解如何解决一些会引发全球灾难的问题。

当然，当我们第一次面对“大数据”这个词时，我们往往会将大部分注意力放在“大”上，并会自动联想起巨人的形象。然而事实上，大数据更多的与大量或者数不清相关。大数据这个词是由前 Meta 公司（后被高德纳收购）的分析师道格莱尼于 2001 年创造出并随之得到广泛传播的，主要用来从三个方面识别数据迅速扩张过程中产生的问题和机会，这三个方面分别包括数据数量、输入/输出速度及数据的多样性[4]。大数据这个概念之所以能在 2000 年以来引起如此广泛的影响主要在于同时代互联网技术的迅猛发展，以及随之产生的海量数据。这种分析大量数据并将其转换成有用信息的重要性已经不言而喻，接下来就需要向数据的三个维度分别进行赋值。如果数据是以信息

流的方式实时传输的，包括一些类似文本、图片或视频的非结构化数据，整合这些不同形式的数据并创造出价值显得尤为重要。因此，数据储备量起到关键作用，然而储备库大小相比之下却无足轻重。研究人员并不需要单纯的数据，他们需要的是数据背后蕴涵的信息。大数据是可以代表大规模数据，然而本质上，分析并捕捉到有价值的数据更为重要。

作为大数据本身，数据的数量一定要足够大。虽然实际运用中并未对数据规模进行特定的规格要求，但通常来看，小规模的数据集会有几万亿字节，大规模的数据集则会达到千万亿级别。表 1.1 里汇集了当前常用的衡量数据大小的单位名称，包括拍字节（PetaByte，PB）、艾字节（ExaByte，EB）、泽字节（ZettaByte，ZB）、尧字节（YottaByte，YB）、Geop 字节（GeopByte，GpB）[5]。以华盛顿国会图书馆为例，其所有馆藏书籍包含的数据合计会有 15 TB。整个 2012 年，人类就累积了 1.27 ZB 的数据。然而，事实上 1 GpB 的数据意味着人类很难彻底了解并且意味着在这些数据基础上还会产生新的数据。

表 1.1　数据大小

数　据	大　小		等　价				
Bit(b) Byte(B) KiloByte(KB)	1 b 8 b 1 024 B	1 2^3 2^{10}	二进制数据（1 或 0） 英文字符（1 字节） 1 页				基本数据单元 一页书 1 200 个字符
MegaByte (MB)	1 024 kB	2^{20}	873	页数	4	书本	一张数字照片：3MB 一首 MP3 歌曲：4MB
GigaByte(GB)	1 024 MB	2^{30}	894 784 341	页数 数字 图片	4 473 256	书本 MP3 音频文件	一到二小时 电影：1~2 GB
TeraByte(TB)	1 024 GB	2^{40}	916 259 689 349 525	页数 数字 图片	4 581 298 262 144	书本 MP3 音频文件	国会图书馆中所有书本的容量 15 TB
			1 613 40	蓝光 CD	233	DVDs	
PetaByte(PB)	1 024TB	2^{50}	938 249 922 368	页数	4 691 249 611	书本	Google 每小时处理的数据量：1 PB
			357 913 941	数字 图片	268 435 456	MP3 音频文件	
			1 651 910 41 943	蓝光 CD	239 400	DVDs	

（续表）

数　据	大　小		等　价				
ExaByte(EB)	1 024 PB	2^{60}	960 767 920 505 705 366 503 875 925	页数 数字图片	4 803 839 602 528 274 877 906 944	书本 MP3 音频文件	美国按周发行的 1 亿份报纸的数据量
			1 691 556 350 42 949 672	蓝光 CD	245 146 535	DVDs	
ZettaByte(ZB)	1 024 EB	2^{70}	983 826 350 597 842 752 375 299 968 947 541	页数 数字图片	4 919 131 752 989 213 281 474 976 710 656	书本 MP3 音频文件	直到 2012 年总的数据量：1.27 ZB
			1 732 153 702 834 43 980 465 111	蓝光 CD	251 030 052 003	DVDs	
YottaByte(YB)	1 024 ZB	2^{80}	1 007 438 153 012 190 978 921 3 843 307 168 202 282 325	页数 数字图片	5 037 190 915 060 954 894 288 230 376 151 711 744	书本 MP3 音频文件	在高速宽带下需要花费 11 万亿年去下载 1YB 的数据
			1 773 725 391 702 841 45 035 996 273 704	蓝光 CD	257 054 773 251 740	DVDs	
BrontoByte (BB)	1 024 YB	2^{90}	1 031 616 699 404 483 562 415 936 393 530 540 239 137 101 141	页数 数字图片	5 158 083 497 022 417 812 079 295 147 905 179 352 825 856	书本 MP3 音频文件	考虑全世界范围内物联网设备实时采集的数据
			1 816 294 801 103 709 697 46 116 860 184 273 879	蓝光 CD	263 224 087 809 782 414	DVDs	
GeopByte (GpB)	1 024 BB	2^{100}	1 056 375 500 190 191 167 913 919 337 402 975 273 204 876 391 568 725	页数 数字图片	5 281 877 500 950 955 839 569 596 302 231 454 903 657 293 676 544	书本 MP3 音频文件	人类能够理解的最大数据量
			1 859 885 876 330 198 730 217 47 223 664 828 696 452 136	蓝光 CD	269 541 465 917 217 192 562	DVDs	

对大数据理解的另一个方面就是数据传输速度及累积率。二十年前的互联网不管是载入高速的数据通信网还是维持日常的基本费用都非常昂贵。然

而，互联网发展到今天，在家、在办公室甚至在大街上，我们都可以使用有线或无线网络轻轻松松地进行百兆级数据的传输。然而，与此同时，数据也在以同样惊人的速度不断地产生和传输。最近几年可以发现，一些自然灾难报道及其他各种爆炸性新闻都是首先通过微博发出的。此外，目前制造型企业使用的智能仪表，家用电器里的智能电视和智能冰箱，无人驾驶汽车都是通过互联网实现巨大数据的实时传输的，随着相关技术的不断成熟，这种数据传输速度还在加快。

对大数据的理解不能仅仅停留于这些种类繁多并持续累积的庞大信息量的表象。之前人们作业的大部分数据都是高度标准化的，管理起来也简单许多。这些数据均被转换成特定的格式且排列整齐，最终形成标准化数据。例如，企业生产过程中的销售收入、存货或故障率都是以往数据分析中常见的且比较容易获得的数据种类。然而，一些新型的数据种类却无法用目前的数据格式去分类，自然无法转化为结构化数据。例如，视频、音乐、图片、地理位置及文本，还有其他很多种类的数据就无法匹配惯用的数据格式，这些都是非结构化数据。这些形式的数据有着不同的大小和内容，很难进行排序。现如今，这些新型数据在以爆发性的速度不断增加，针对这些数据的处理方式也亟待问世。

考虑到大数据的数据规模、传输速度和数据种类，通常一个数据集会达到数千太字节。数据的处理和传输速度从几秒到数小时不等，并且这种数据能以任何一种结构化或非结构化的形式存在，这就给目前惯用的数据处理和分析方式提出了革命性的挑战。更进一步说，大数据虽蕴涵大量不同种类的数据，但其精髓仍然在于如何有效地对数据进行分析以提炼出有意义的信息。因此，大数据既包括那些用现有技术方法无法有效管理和解读的大量数据，还包括管理信息、分析信息的人力资源、组织机构和相关技术，这些都构成了一个完整的大数据。通过这个管理和分析过程，最后得出的结论正是大数据分析的价值所在。

如今，社会变得更加互联互通，数据也呈爆发性产生。但这并不意味着只要数据集足够大，包含数据足够多，就值得我们去分析，去得出一些结论。数据大量增加的同时意味着一些不重要的或没有意义的数据也掺杂在其中并随之增加，此时就需要分析人员独具慧眼，剔除这些无效的原始数据，过滤出有分析意义的数据集。大数据注重开放的思维和广阔的视角，它不仅关乎数据数量、传输速度或数据本身的大小，还关乎一个巧妙的视角和对趋势的预测。我们不必离开“森林”，相反，我们可以登上“山顶”。同样，为能深入理解大数据，我们必须脱离现有的思维和视角，正所谓“站得高方能看得远”；为看到从内部看不到的风景，我们需要一个不一样的视角；为看到更多，我们必须依靠大数据。对大数据的分析会大幅提升我们的视野。

1.2　是什么产生了大数据

从互联网兴起到 2012 年的统计数据显示，我们已经累积了 1.27 ZB 的数据，并且在 2016 年，全球网络数据通信量超过 1ZB[6]。我们无法真正了解如此大规模的数据流是如何产生的，其中一个原因在于数据存储工具的不断升级换代催生了数据的爆发式累积。随着人类文字的诞生，早先刻在树皮、动物皮、树枝（如竹子）、石头、黏土碑等上面的历史记录，由于现代打印技术的革命性发展开始以纸张的形式记录下来。过去因为缺乏必要的记录轨迹，许多人类活动和知识经验都无法保留下来，而现在却可以利用文字完整地记录在纸张上。然而，纸张保存信息也有弊端，往往纸张体积增加的速度远远大于其内含信息量增加的速度。进入 20 世纪，随着模拟存储设备的诞生，人类活动的信息记录载体开始变为体积更小的胶卷或磁带，如胶片、照片、磁带盒、录像带等，这些工具体积虽小但包含的信息量却远超传统的纸张。在数字化时代来临前的 20 世纪 80 年代，人类已经累积了 2 620 000 TB 的数据，90%以上都是以胶片或磁带的形式保存的[7]。而 1990 年数字化革命的到来，这些记载工具又进行了升级，文字、语音、图片、影像都开始被数字化，数据存储能力也显著提升。最初的计算机存储工具叫作软盘，之后出现了硬盘和移动闪存 U 盘。如今，我们可以在智能手机上保存数十 GB 的数据，还能随时随地查阅电子书、图片、音乐和影像。如果把人类一直以来累积的数据保存在 CD 盘上，然后一张接一张地摞起来，其高度相当于从地球到月球距离的 6 倍。科学技术的进步也大大降低了存储数据的成本。1980 年一块 1 GB 的硬盘就要花费 21 万美元，而到了 2013 年，这一金额仅仅为 3 美分。成本的大幅缩减自然大大促进了数据累积的迅速增加（见图 1.1）。