“十二五”国家重点图书出版规划项目

大数据技术与应用

丛书策划

上海大数据产业技术创新战略联盟（上海产业技术研究院）

上海市数据科学重点实验室（复旦大学）

丛书主编

朱扬勇 吴俊伟

Big Data

Technology and Application Series

蔡立志 武 星 刘振宇

主编

大数据测评

上海科学技术出版社

大数据技术与应用

大数据测评

蔡立志　武　星　刘振宇
主编

上海科学技术出版社

本书出版由上海科技专著出版资金资助

图书在版编目(CIP)数据

大数据测评 / 蔡立志，武星，刘振宇主编. —上海：
上海科学技术出版社，2015.1(2015.1 重印)
(大数据技术与应用)
ISBN 978-7-5478-2278-4

Ⅰ.①大… Ⅱ.①蔡… ②武… ③刘… Ⅲ.①数据处理—评价—研究 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2014)第 133592 号

大数据测评
蔡立志 武 星 刘振宇 主编

上海世纪出版股份有限公司
上海科学技术出版社 出版
(上海钦州南路 71 号 邮政编码 200235)
上海世纪出版股份有限公司发行中心发行
200001 上海福建中路 193 号 www.ewen.co
苏州望电印刷有限公司印刷
开本 787×1092 1/16 印张 13.25
字数：300 千字
2015 年 1 月第 1 版 2015 年 1 月第 2 次印刷
ISBN 978-7-5478-2278-4/TP·27
定价：52.00 元

内容提要

大数据技术的发展，在带来产业快速发展的同时，也带来了很多软件技术的新需求。本书介绍了大数据的概念和特征，各国大数据的发展战略、发展趋势及其标准化情况，以及对软件测试带来的挑战。在此基础上，对面向大数据处理框架、大数据基础算法、应用系统、系统安全和隐私泄露等测评技术展开了分析和讨论。在底层支撑框架层聚焦于单元测试和框架基准测试；在基本算法中涵盖了聚类、分类及其个性化推荐；在应用层，介绍了其性能测试中若干问题，重点阐述数据集的设计与分析；在全书的最后，讨论了大数据的安全和隐私问题，突出介绍由于大数据所引发的新安全问题及其对策。

本书综合了众多业界专家、作者、学者的研究和产业成果，通过对大量文献和材料分析编著形成，可为从事大数据或者软件测评的学者、软件工程研究人员、高校研究生、大数据产业人员提供参考。

大数据技术与应用

学术顾问

大数据技术与应用

编撰委员会

本书编委会

主　编

上海计算机软件技术开发中心　**蔡立志**
上海市计算机软件评测重点实验室
上海大学　**武　星**
上海计算机软件技术开发中心　**刘振宇**
上海市计算机软件评测重点实验室

编　委

上海计算机软件技术开发中心　**陈敏刚**
上海市计算机软件评测重点实验室
上海计算机软件技术开发中心　**陈文捷**
上海市计算机软件评测重点实验室
上海计算机软件技术开发中心　**曹祥琼**
上海市计算机软件评测重点实验室
上海市公安局网络安全保卫总队　**沈与辛**
华东理工大学　**郑　阳**
中国电子技术标准化研究院　**王洁萍**
上海计算机软件技术开发中心　**吴建华**
上海市计算机软件评测重点实验室
上海计算机软件技术开发中心　**胡　芸**
上海市计算机软件评测重点实验室
浙江省电子信息产品检验所　**宣以广**

序（一）

软件质量具有功能性、可靠性、易用性、效率性、维护性和可移植性六个特性，可以从软件的内部质量、外部质量和使用质量三个视角去考量。软件质量保证就是把质量嵌入到软件生存周期全过程中，以保证软件的“生产”质量。而软件测评是软件质量保证的一个关键手段，也是软件产品发布前的最终检验，目前其技术和工具亦日趋成熟。但是，今天已经面临“大数据”时代，带来的挑战是不言而喻的，这将是软件工程领域一次重大转折，由关注程序和产品为中心转向关注数据和服务为中心，从而其质量保证将会有全新的面貌。国际标准化组织和国际电工委员会第一联合技术委员会（ISO/IEC JTC 1）启动了数据质量和信息技术服务质量的标准项目，这充分说明我们应当重视这种重大的变化。

大数据的 4V 特性带来软件测试新的挑战：输入集的构建正面临着新的变化，输入不是一个数据或者几个数据，而是一个庞大的数据集；输入数据的样本覆盖和实际应用的匹配度如何；数据量是否能够满足关于数据相关性分析的要求；数据类型包括结构化数据和非结构化数据；输出集的正确性判断也面临新问题；输出结果的不确定性带来软件测试的 ORACLE 问题等。

今年春节，我和蔡立志博士他们闲聊时，从微信话题开始逐步转移到另外一个热门话题：大数据的技术及其发展。我问小蔡对这些挑战，有什么新的软件测试方法和技术，从而获悉了蔡立志博士所带领的团队正在编写《大数据测评》一书。该书有以下特点：

（1）贴近标准　概念和术语来自标准，而且还介绍了大数据基础与技术标准体系框架；当然，标准的制定需要研究的成果，尚需时日。

（2）资料翔实　只有占据了大量资料才有发言权，作者采用大数据的概念方法来阐说大数据的应用与测评。

（3）结构清晰　对支撑架构、算法、应用性能和安全等方面的测评方法和技术都做了深入的介绍，有利于应用。

作为一名长期从事软件工程研究的工作人员，有幸能和这些年轻人一起探讨，能够感悟到他们对生活的向往和激情，也使自己的心态年轻。同时，我亦欣赏他们直面问题、勇于创新的精神；欣赏他们积极进取、主动挑战的理念；欣赏他们不断努力、务实工作的态度。在这个高速发展的互联网时代，太需要更多的年轻人的创造性工作，并在不断实践基础上主动地“扬弃”，来解决我国信息化发展中的新问题。该书可能无法覆盖大数据测评所有问题，相信方法和技术会有更大的进步，后来者居上是作者的期望。该书的出版，将会有益于读者掌握一门重要的技术手段，有益于大数据技术应用的普及，有益于我国信息产业的发展。对大数据开展测评，具有较强的实现意义和应用价值，本书是一本值得推荐的书籍。

朱三元

2014 年 5 月 4 日

序(二)

大数据技术正在深刻地影响着社会的方方面面。从早上起来查看天气预报、食用营养早餐、出行查询交通导航、网上购物的个性化推荐等,无不体现着大数据对人们生活的影响。

2012 年中国软件测评机构联盟技术委员会开会时,技术委员会主任蔡立志博士就大数据问题和我做过交流,但是并没有形成特别清晰的思路。2013 年 3 月中国软件测评机构联盟在杭州召开了技术委员会会议,蔡立志博士在会上作了“大数据对于软件测试的挑战”的学术报告,使得全国近 50 家的联盟成员单位分享了他的思考成果。在年轻同志的努力下,我国的软件测试技术和产业最新技术能够同步发展,使联盟的工作取得非常可喜的成效,我感到特别欣慰。

关于大数据的测评问题,存在两种极端的思维模式:一是大数据软件也是一种软件,没有什么特别的技术挑战;二是大数据软件由于其输入和输出的复杂性,根本无法测试。在中国软件测评机构联盟技术委员会的学术交流上,这两种观点都有较多的支持和拥护者,争论得比较激烈。我个人觉得这两种思维都存在一定偏颇,既要兼顾技术发展的新特性,又不能因为其复杂性而不去探索,否则测评技术将永远无法跟上产业发展的步伐。在听了“大数据对于软件测试的挑战”的报告以后,我建议蔡立志博士将思考也可以说是初步的研究成果做进一步的深化和沉淀。

另一方面,作为第三方软件测评机构,其基本的公信力是建立在测试步骤的一致性、测试结果的重现性、立场的客观性之上。在产业无法达成共识,又没有统一的标准时,论文和书籍的编写可以很好地弥补第三方测评在这方面的不足,为测评人员提供技术指导和思路。而现在大数据就面临着这一现状,急需本领域前瞻性的技术指导,这也是最近几次和同行交流达成共识。

上个月突然接到蔡立志博士的电话，请我对其团队编写的《大数据测评》做一个序，我一口气读完了全书。这本书涵盖了大数据分析框架测试、算法质量测试、性能测试、大数据安全和隐私各个方面。内容翔实，覆盖面广，操作性强，可以为各个大数据的研究和测评技术人员提供有价值的参考。作为一个在软件质量和标准化领域工作近40年的老同志，我希望本书编写团队做进一步努力，努力将其转化为国家标准或行业规范。这是我个人的希望，也是产业的希望，我相信这个时间应该不会太远。

中国软件测评机构联盟秘书长

冯　惠

2014年4月13日

前 言

在软件测试的经典定义中，这样描述软件测试“为发现软件错误，而运行软件的活动”。其基本的思路是根据软件需求规格说明书，执行软件操作和输入数据，依据软件实际输出结果和预期输出结果来评判软件是否满足规定的要求。

单元测试，要求依据软件实现的内部结构编写各种测试用例。语句覆盖、条件覆盖、判定覆盖、路径覆盖等覆盖准则的一个基本前提就是能够对软件的执行逻辑进行正确分析。随着各种大数据处理 PAAS 平台（Platform-as-a-Service，平台即服务）的出现，这种情况也在发生新的变化。测试人员看不到完整的逻辑，而是中间一部分，单元测试如何做？如果软件运行在分布式集群中，单元测试中的覆盖如何实现？大数据应用处理的不是静态的数据，同时大数据开放性数据的来源、数据的质量、数据的类型也并不严格受软件所控制。

2005 年一个纯属偶然的机会，有几个用户要求上海市计算机软件评测重点实验室测试和评价类似“热度识别”、“趋势分析”等软件。这类软件的共同特征就是不具备类似“1＋1＝2”特性：软件输入不是一个特定含义数据，而是源源不断输入的数据集，例如论坛、新闻评论、博客等；软件输出没有客观的正确性的判断条件。例如，一篇关于讨论汽车企业上市的新闻，到底归属于哪一类，证券类还是汽车类，不同人由于其关注度不同导致了同样的对象得出不同结论。在热点识别时，不同的人在讨论同一件事不会完全采用同一词语、同一语句，必须采用某种相似性判定函数，如余弦相似性计算函数，对给定的两篇信息做出判断，即它们是否讨论同一个事件，而相似不是一个确定的概念，而是一个模糊的概念。在趋势分析时，没有一套趋势曲线和实际发展曲线完全重合的，意味着对软件系统的评判只有优劣之分，而没有对错之分。

大数据分析是一把双刃剑，在分析数据中存在的价值的同时，会带来新的隐私泄露途

径和手段。这些隐私泄露的途径与手段和其他信息安全问题存在很大的不同，具有很强的隐蔽性。分析发布的数据，必须注意是否在不留意的过程中将隐私信息也发布了。

这些新的测试技术需求一直萦绕在我的脑中很多年，也没有特别好的解决方案。2012 年左右，产业开始出现了火热技术趋势“大数据”，回想纠结这么多年的测试需求，就是由于“大数据”的 4V 特性所形成，我们开始关注搜集关于大数据测试的相关技术，包括底层支撑的分布式处理框架、典型的算法，以及产生的隐私泄露问题。2012～2013 年，在中国软件测评机构联盟的多次技术交流会议中，我把关于这方面的技术思考做了交流，不断得到了同行们的支持和鼓励，技术思路也逐渐变得清晰。2013 年在上海大数据产业技术联盟的倡议和支持下，决定把这些想法编著成书，以便和同行们分享交流。

针对上述的新问题、新需求，本书以 Hadoop 为主线开展大数据测评的探讨。在底层支撑框架层聚焦于单元测试和框架基准测试；在基本算法中涵盖了聚类、分类及其个性化推荐；在应用层，介绍了其性能测试中若干问题，重点阐述数据集的设计与分析；在全书的最后，讨论了大数据的安全和隐私问题，突出介绍由于大数据所引发的新安全问题及其对策。在本书的编著过程中，得到上海计算机软件技术开发中心、中国电子信息标准化研究院、上海微趣信息技术有限公司等单位在时间、人员、技术等多方面的大力支持。感谢网宣办的徐良奇老师，每次和徐老师关于具有类似大数据特征的软件测评讨论，都让我受益匪浅，激发了我对于这方面问题思考的动力。大数据各项技术处于快速的发展过程中，所涉及的范围也十分庞大，本书选择了大数据测评技术中几个相对较为成熟的点，并未覆盖所有技术点。在本书的编著过程中，收集了大量的文献资料，包括最新的网页信息，本书的编著离不开这些宝贵的资料，在此一并表示感谢。限于作者的水平，书中肯定有不足和遗漏，任何的意见和建议，请发送电子邮件：clz@ssc.stn.sh.cn。

蔡立志

目 录

第1章 绪论 1

• 1.1 概述 2
• 1.2 大数据战略与趋势 6
1.2.1 大数据战略 6
1.2.2 大数据趋势 8
• 1.3 大数据标准化研究 12
1.3.1 国外标准发展现状 12
1.3.2 国内标准发展现状 14
• 1.4 大数据应用 16
1.4.1 趋势预测 17
1.4.2 疫情分析 17
1.4.3 消费行为分析 18
1.4.4 智慧金融 20
1.4.5 精确营销 20
1.4.6 舆情分析 21
• 1.5 大数据对软件测试的挑战 23
参考文献 24

第2章 面向大数据框架的测评 27

- 2.1 概述 28
- 2.2 面向数据质量的测评 29
- 2.2.1 数据质量 29
- 2.2.2 数据预处理 31
- 2.2.3 数据质量测评 36
- 2.3 分布式数据模型及测试 40
- 2.3.1 框架 40
- 2.3.2 数据模型 41
- 2.3.3 单元测试 43
- 2.4 大数据的基准测试 48
- 2.4.1 基准测试 48
- 2.4.2 测试方法 48
- 2.4.3 测试内容 50

参考文献 63

第3章 大数据智能算法及测评技术 65

- 3.1 概述 66
- 3.2 聚类算法及测评 67
- 3.2.1 聚类及其在大数据中的应用 67
- 3.2.2 聚类的典型算法及分析 68
- 3.2.3 聚类算法的测试 72
- 3.2.4 聚类质量的评估 76
- 3.3 分类算法及评估 79
- 3.3.1 分类及其在大数据中的应用 79
- 3.3.2 分类的典型算法及分析 80
- 3.3.3 分类算法的测试 86
- 3.3.4 分类器性能的评估 88
- 3.4 推荐系统算法及其测评 92
- 3.4.1 推荐系统算法 94

3.4.2 推荐系统的测评实验 97
3.4.3 推荐系统的评估 99
参考文献 104

第4章 大数据应用的性能测评技术 107

·4.1 **概述** 108
·4.2 **大数据应用的影响因素与性能测评** 109
4.2.1 影响大数据应用的因素 109
4.2.2 大数据应用的性能测评类型 109
4.2.3 大数据应用的性能测评指标 110
·4.3 **大数据应用测试的支撑数据设计** 113
4.3.1 大数据的数据结构特点 113
4.3.2 大数据的数据设计依据 114
4.3.3 大数据的数据生成方法 116
·4.4 **大数据应用性能测评模型** 117
4.4.1 应用负载模型 117
4.4.2 数据负载模型 122
·4.5 **工具与案例** 130
4.5.1 性能测试工具 130
4.5.2 性能测试流程 131
4.5.3 某网络舆情监测系统测试案例 134
4.5.4 某微博大数据平台测试案例 137
参考文献 139

第5章 大数据应用的安全测评技术 143

·5.1 **概述** 144
·5.2 **影响大数据应用安全的要素** 145
5.2.1 影响架构安全的要素 145
5.2.2 影响数据安全的要素 148
·5.3 **大数据架构的安全测评** 150

5.3.1 分布式计算框架的安全测评 150
5.3.2 非关系型数据库的安全测评 155
·5.4 **数据的安全性测评** 160
5.4.1 数据来源的安全性测评 160
5.4.2 隐私保护程度的测评 164
·5.5 **应用安全等级保护测评** 175
5.5.1 用户鉴别 176
5.5.2 事件审计 177
5.5.3 资源审计 179
5.5.4 通信安全 181
5.5.5 软件容错 182
参考文献 182

索引 185