课程 | 实验 | 题库

教育部产学合作协同育人项目成果教材
西普教育研究院 IT 前沿技术方向高校系列教材

Big Data Testing Technology

大数据测试技术

数据采集、分析与测试实践

在线实验 + 在线自测

刘攀 主编
王海龙 徐振宁 时允田 林雪纲 副主编

实验吧
让实验更简单

-SaaS
开放实验云平台

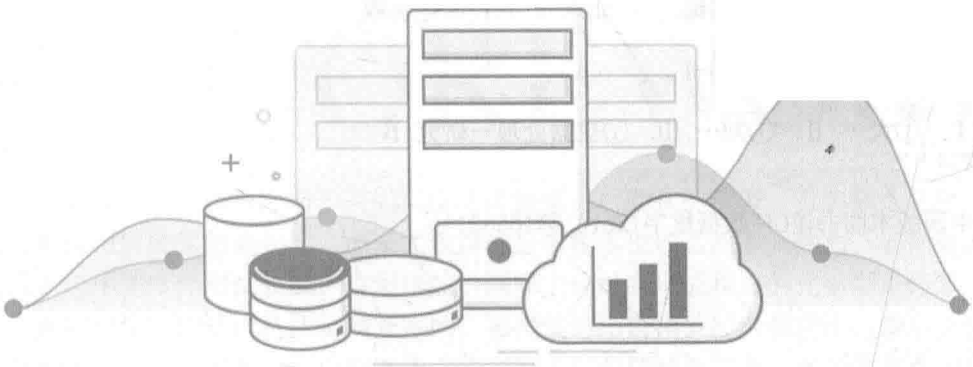
- 内容新颖，可操作性强，层层深入，简明易懂。
- 从实用角度出发，重点培养动手解决问题的能力。
- 提供体系完整的在线实验，即学即练，书网结合。

扫描书中二维码随时进行在线测试

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

-SaaS 课程 | 实验 | 题库

教育部产学合作协同育人项目成果教材
西普教育研究院 IT 前沿技术方向高校系列教材



Big Data Testing Technology

大数据测试技术

数据采集、分析与测试实践

在线实验 + 在线自测

刘攀 主编
王海龙 徐振宁 时允田 林雪纲 副主编

人民邮电出版社
北京

图书在版编目（CIP）数据

大数据测试技术 ：数据采集、分析与测试实践 ：在线实验+在线自测 / 刘攀主编. -- 北京 ：人民邮电出版社，2018.9
ISBN 978-7-115-48953-1

Ⅰ. ①大… Ⅱ. ①刘… Ⅲ. ①数据处理－研究 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第165851号

内 容 提 要

本书从工程角度出发，阐述了运用大数据技术进行软件测试的方法和实现过程。全书共 11 章，介绍了大数据测试思维、手机联网数据的收集方法、数据格式的转换方法、大数据的处理方法、软件缺陷挖掘技术及实践项目的应用等。本书第 1 章介绍了大数据的测试方法和思维方式，随后每一章都通过案例来讲解大数据技术的相关理论及其测试应用。除第 1 章外，每章最后都提供了思考题来帮助读者回顾和巩固本章的学习内容，本书最后还提供了思考题的参考答案。同时，本书对每一个案例进行了详细图例展示和讲解。

本书可以作为应用型本科、高职高专计算机相关专业大数据技术及测试相关课程的教材，也可以作为大数据测试技术培训教材。本书同样适合作为金融、证券、保险和与数据分析相关行业从业人员的自学指导书。

◆ 主　　编　刘　攀
副 主 编　王海龙　徐振宁　时允田　林雪纲
责任编辑　左仲海
责任印制　马振武

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn

固安县铭成印刷有限公司印刷

◆ 开本：787×1092　1/16
印张：16.25　　2018 年 9 月第 1 版
字数：376 千字　　2018 年 9 月河北第 1 次印刷

定价：49.80 元

读者服务热线：(010)81055256　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

前 言 FOREWORD

在大数据时代到来之前，人们很少关注数据背后潜在的价值。直到最近几年，人们逐渐认识到数据的重要性，面向大数据的分析技术也应运而生。然而，如何将大数据技术应用于软件测试领域，直到目前，国际上也没有形成相关的完整理论体系。通常，人们对事物的认识是从实践开始的，再慢慢总结出相关经验和理论，最后又依照理论来指导实践。因此，在构建基于大数据的软件测试理论之前，必须先在实践中进行大量的大数据测试应用。

编者在攻读博士学位之前，一直在软件公司从事开发工作，读博之后开展了一些有关软件测试理论方面的研究。2012 年后，大数据这个概念逐渐被人们接受，编者也十分留意学术界及产业界有关大数据测试的动态。然而，除了中科院等少数单位外，国内很少有单位进行大数据测试的相关研究，市面上也仅有数量不多的针对大数据系统的测试工具。编者也一直在思考：大数据测试难道仅仅指对大数据系统的测试吗？这与原有的软件测试方法又有多少不同呢？难道不能将大数据的技术应用到软件测试中吗？为此，编者也对一些研究院所和企业进行了调研，希望能够发现将大数据技术应用到软件测试中的方法。直到 2017 年，在产业界朋友的帮助下，编者首次完整实现了一套由大数据收集、大数据分析、软件缺陷挖掘组成的大数据测试实践操作方法，并成功应用于手机联网数据分析和 Web 日志数据分析。这套操作方法并不需要用户去购买额外的软件，仅仅使用市面上的一些共享软件就能实现，因此，普通用户也可以在自己的计算机上搭建相关的环境，实现大数据测试。

适逢教育部协同育人项目及上海市二类高原学科（应用经济学科商务经济方向）资助，编者才开始着手编写大数据测试技术的实践操作方法。在此基础上，编者对这些实践操作进行了总结，最终形成了这本书。希望这本书能够帮助人们转变软件测试的传统思维模式，即不再满足于从技术角度出发设计测试用例以实现软件测试，而从用户使用的角度出发，依据用户使用的数据探测软件中的缺陷，同时也希望能将大数据测试思想、方法和实践操作方式推广到工业应用中。

上海商学院的刘攀副教授编写了本书的第 1~4 章和第 8~11 章，美国南缅因大学商学院（School of Business at the University of Southern Maine）的徐振宁（Zhenning (Jimmy) Xu）教授编写了第 5~7 章。上海钒兆钛智能科技有限公司联合创始人兼首席客户服务官 CCO 王海龙为本书提供了大量的数据资源，北京西普阳光教育科技股份有限公司的时允田和林雪纲为本书提供了在线实验和在线测试平台的技术支持。

由于编者才疏学浅，书中必有疏漏及不足之处，还望读者批评指正。愿这本书能够成为读者进入大数据测试领域的一块垫脚石。

编者

2018 年 1 月

平台支撑 PLATFORM SUPPORT

为了让广大学习者能够快速入门，本书以实践案例为主线，通过遵循书中案例的操作步骤，完成一个个实验案例，来学习大数据测试技术。同时，北京西普阳光教育科技股份有限公司（简称西普教育）开发的在线教育平台——实验吧（http://www.shiyanbar.com），提供了强大的集成实验环境及海量的在线教学资源，把配套的实验搬到线上，可以让读者更方便地结合本书进行实践。

1. 如何学习本书中的配套实验课程

（1）购买本书后，找到粘贴在本书封底的刮刮卡，刮开并获得学号。

（2）登录实验吧网站（www.shiyanbar.com），完成网站注册。

（3）登录人邮学院在线实验中心（rymooc.shiyanbar.com），输入在实验吧注册的账户及密码，完成登录（见图 1）。

（4）输入刮刮卡中的学号、姓名填写“人邮学院”，单击“保存”按钮，完成绑定（见图 2）。

（5）完成绑定后，自动登录进入在线实验中心，开始学习本书配套的课程资源。

图 1 登录在线实验平台

图 2 绑定学生信息

2. 如何学习本书中配套练习题

实验吧教研团队为本书配套了丰富的课后练习题，读者通过扫描本书各项目里配套的习题二维码，即可进行在线自测，提交后自动判断正误，并提供正确答案。

目 录 CONTENTS

第 1 章 大数据测试的思维

为了保证软件的可靠性，人们通常需要采用一些技术（如白盒测试和黑盒测试）来设计测试用例并测试软件。实践证明，这些技术是非常必要的。随着大数据技术的发展，软件从业人员也在尝试着采用大数据技术来保证软件的可靠性。通常，软件在发布之前已经经过了若干次测试，但发布后仍可能存在缺陷。如何探测这些软件缺陷呢？传统方式是采用更为严格的测试方法，设计更多的测试用例并测试软件。事实上，软件在发布后已经被众多的用户使用，生成了海量的用户使用数据（包含软件出错数据），如果能对这些数据进行分析，就能探测出软件中的缺陷。这种测试方法并不需要测试人员设计测试用例，而是采用海量用户使用数据作为测试用例；这种测试方法也不需要运行测试用例，因为用户的使用过程就是软件测试的过程。因此，这是一种全新的软件测试方法，必将带来软件测试思想的变革。本章将重点介绍大数据背后的规律及其在软件测试中的应用。

1.1 规律是否存在

今天人们正面临着急速的数据膨胀，2003 年人类创造的数据还仅有 10^{18} 个字节，到 2012 年，数据已多达 10^{21} 个字节，2015 年的全球数据是 2012 年数据的 8 倍，而预计到 2020 年，全球数据将达到 44ZB 个字节（1ZB=10 万亿亿字节）。事实上，大数据时代已经来临，基于大数据分析的各种应用正改变着人们的社会、生活和工作，也为企业带来了新的商业发展机遇。例如 2012 年 12 月 12 日，淘宝网推出了“时光机”，根据淘宝买家几年来的商品购买记录、浏览点击次数、收货地址等数据编辑制作了“个人网购日志”。该日志就是基于对 4.7 亿淘宝注册用户网购数据的分析实现的，是一个经典的大数据应用。

目前，越来越多的公司将数据当作一种重要的战略资源，进行数据储备和数据分析。与以往不同，在大数据时代，人们对数据的认识和处理方式发生了新的转变。过去，人们将数据看作是静止的、陈旧的，数据之间是无关联的，对数据的处理仅限于简单的查询和分类统计，并以此得出一些人们自认为存在的规律。**然而，事实上，当人们对海量数据进行分析之后，往往会发现某些规律根本不存在。**

例如对一个物流仓库数据进行收集，希望发现并预测仓库中的哪些商品会延期交货，在开始阶段收集到的需要延期交货的数据如表 1.1 所示。

根据以往的经验，人们通常认为库存为零，或者库存较少，但未来预期销售较多，而运输数目又少的产品极可能出现延期交付现象。而且表 1.1 中的数据也恰恰符合这种认识。

然而，当获得了 10 万条库存记录后，会发现其中许多库存数量大的商品也会延期交付。表 1.2 显示了部分库存量较大且会产生延期交付的产品数据。

表 1.1　开始阶段收集到的需要延期交货的数据

编号	当前库存	运输时间	运输产品数	预期未来 3 个月销售数	预期未来 6 个月销售数	……	是否延期交付
1	0	2	0	0	0	……	是
2	6	8	0	0	0	……	是
3	0	8	0	4	7	……	是
4	7	8	0	11	27	……	是
5	0	8	0	9	14	……	是
6	0	4	0	10	10	……	是
7	0	8	0	2	3	……	是
8	0	8	0	22	44	……	是
9	0	2	0	16	16	……	是
10	3	8	0	25	39	……	是

表 1.2　部分高库存且会产生延期交付的产品数据

编号	当前库存	运输时间	运输产品数	预期未来 3 个月销售数	预期未来 6 个月销售数	……	是否延期交付
164	4197	21000	29400	39900	40710	……	是
6705	232	8	25	1272	2148	……	是
6715	117	2	0	347	543	……	是
6898	224	2	0	2875	5259	……	是
7478	212	8	144	4364	6432	……	是
7534	495	8	286	1601	2441	……	是
7600	545	12	169	2976	5808	……	是
7981	1636	8	396	4035	8192	……	是
9038	10302	2	0	6820	18700	……	是
101773	1085	32	32	32	1	……	是

另外，通过对不需要延期交货的数据进行统计，还会发现大量库存少的产品不会延期交付，如表 1.3 所示。

表 1.3　库存少但不需要延期交付的产品数据

编号	当前库存	运输时间	运输产品数	预期未来 3 个月销售数	预期未来 6 个月销售数	……	是否延期交付
61	0	0	0	0	0	……	否
62	0	0	0	0	0	……	否
63	0	0	0	0	0	……	否
64	0	0	0	0	0	……	否
65	5	2	0	0	0	……	否
66	6	12	0	0	0	……	否
67	1	12	0	0	0	……	否
68	0	2	0	0	0	……	否
69	0	2	0	0	0	……	否
70	12	0	78	168	306	……	否

过去，人们形成的“经验”往往是基于非大规模数据的总结，而当数据规模达到一定数量时，原来的“经验”就变得不一定准确。比如基于大数据分析之后，AlphaGo 曾下出了人类长期以来认为是亏损的棋路。

1.2　大数据的背后

现在，依据大数据的统计和分析能够发现很多以前无法想象的规律。例如，闻名于世的啤酒与尿片的故事就是大数据分析的一个经典案例。全球零售业巨头沃尔玛在对消费者购物行为分析时发现，男性顾客在购买婴儿尿片时常常会顺便搭配几瓶啤酒来犒劳自己，于是尝试推出了将啤酒和尿片摆在一起的促销手段，没想到这个举措居然使尿片和啤酒的销量都大幅增加了。如果转换一种思路来理解这种现象，考虑“奶爸”在世界杯期间既要看足球比赛，又要带自己的孩子，为了不耽误他们观看完整的足球比赛，他们会将孩子放在边上，任由他们玩耍，而自己喝啤酒庆祝球队的胜利，因此，他们需要购买尿片来帮助他们的小宝宝隔尿，同时也不影响自己喝啤酒庆祝球队的胜利。

另一个经典的例子就是 Google 预测流感案例。2009 年 2 月，国际权威期刊 *Nature* 上刊登了一篇名为 *Detecting Influenza Epidemics Using Search Engine Query Data* 的文章，论述了 Google 基于用户的搜索日志（包括搜索关键词、用户搜索频率以及用户 IP 地址等信息）成功“预测”了流感病人的就诊人数。Google 是如何做到的呢？原来美国有个疾控中心，它统计了美国本土各个地区的疾病就诊人数，一般会延迟两周公布数据。而 Google 利用其搜索引擎搭建了一个预测平台，提前获取疾控中心的大数据，并对这些数据进行回归分析，从而成功地预测了复杂的流感规模问题。

在国内，杭州市首次利用大数据治理交通阻塞问题。2016 年 10 月，杭州市政府联合阿里云公布了一项计划：建立杭州市的城市大脑。城市大脑的内核将采用阿里云 ET 人工

智能技术，对整个城市进行全局实时分析，自动调配公共资源，修正城市运行中的问题，并最终进化成为治理城市的超级人工智能。“缓解交通阻塞”是城市大脑的首个尝试，并已经在萧山区的中心路段投入使用，部分路段车辆通行速度提升了 11%。

上述 3 个例子说明，通过对大数据进行分析，可以找到一些隐藏在数据背后的规律。那么，能否将大数据的处理方法运用到实际的软件测试中，以便发现软件中的缺陷呢？**如果能够获得海量的用户使用软件的数据，再利用大数据处理技术对这些数据进行分析，就能从中发现使软件执行失效的小概率事件，从而发现软件缺陷，这种测试方法称为大数据测试方法。**

1.3 大数据测试方法

传统软件测试过程中，由于测试成本的约束，测试用例的总数是小样本的有限集合。但当软件拥有复杂的结构或采用构件化设计时，测试人员很难设计出测试用例来完全覆盖软件中的所有构件组合；当软件处于复杂的使用环境时，测试人员很难设计出测试用例来模拟软件实际使用中的所有场景；当软硬件紧密结合开发时，测试人员很难设计出测试用例来涵盖每一次的软硬件组合。

传统软件开发的生命周期是从需求分析开始，到软件产品发布截止的。但随着时代的发展，软件并非是在发布之后就停止更新，往往会在很短的时间内继续推出新的版本，如部分品牌的手机操作系统已经实现了一月甚至一周一更新。因此，从长时间看，软件处于一个长期迭代的过程，可以在每一次软件开发和修改时采用传统软件测试技术进行测试，并在软件的某一版本发布之后实施大数据测试。

如图 1.1 所示，在经典的瀑布式软件开发流程中，软件测试包含单元测试、集成测试、系统测试和验收测试 4 个阶段。单元测试往往不会过多地涉及其他程序或者模块，因此软件复杂程度对单元测试阶段的影响是最小的。在集成测试和系统测试阶段，软件复杂程度的增加会导致测试成本的激增。但在这两个测试阶段，用户参与程度低，因此很难获得大量用户使用软件时产生的数据。一般在软件的某个版本发布之后，用户才会大量使用该软件，这也为收集用户的使用数据提供了可能。如果能对这些数据进行深入的大数据分析，就能发现传统软件测试阶段不易发现的错误。这种做法的好处有两点：一是节约了测试成本，海量的用户在真实环境中对软件进行了“操作性测试”，而且这个过程并不需要测试人员去构造和执行测试用例；二是大数据测试方法能够发现传统软件测试阶段难以发现的软件错误。如在实际使用中，软件的操作环境是极为复杂的，而在传统软件测试阶段，测试人员很难模拟所有测试场景。又比如在对服务器的压力测试中，用户对软件的某种极端操作会导致服务器不定期产生错误（小概率事件），而这些错误很难在传统的压力测试环节暴露出来。只有通过收集海量用户使用软件的数据，使用大数据测试方法才能有效地发现这种小概率事件。

图 1.2 显示了大数据测试方法的测试流程。传统软件的 4 个测试阶段是在软件版本发布之前完成的，**大数据测试方法包括用户使用、数据收集、大数据分析及缺陷挖掘 4 个阶段。**与敏捷开发方法类似，大数据测试方法同样需要用户的参与，不同的是参与用户的数量。敏捷开发以用户的需求进化为核心，采用迭代、循序渐进的方法进行软件开发，因此

仅需少量的用户参与开发，提出并修改需求。而大数据测试方法需要收集海量的用户使用数据，这相当于海量用户对软件进行测试。此外，如何收集海量用户使用数据则是大数据测试的另一关键技术。传统用户使用数据的收集一般都是被动的，即软件提示用户书写并发送软件崩溃时的相关信息给软件公司。这种方式原始又落后，因为大多数用户会因为懒惰而不去做这件事情。大数据时代用户使用数据的收集是主动式的，即自动模拟用户的操作，并采用软件自动抓取用户操作失效时的数据，再在大数据分析阶段利用大数据的一些相关技术来处理并分析海量的数据。在传统关系数据库中，同一字段属性中有且仅有同一类型的数据，属性中的每一个数据都是一个不可分割的项。但大数据则不同，同一属性中的数据可以是不同的类型，也可以由多个项构成，因此，传统数据库的处理方法不能对大数据进行有效处理，需要运用专门的大数据处理技术。软件缺陷挖掘是指从海量数据中发现一个软件 Bug 后，采用大数据技术再次从数据中挖掘出更多具有相同特征的软件 Bug。

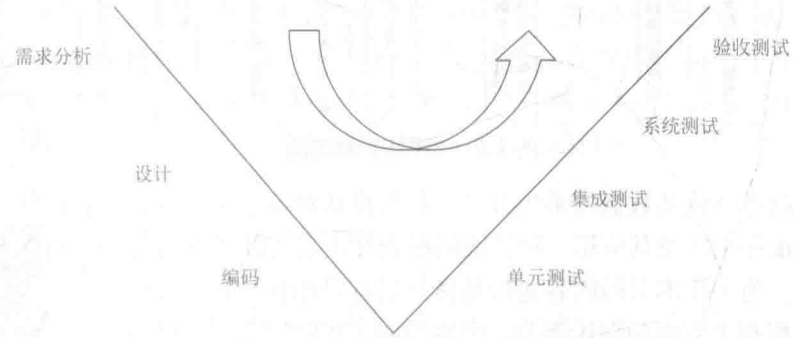


图 1.1　在瀑布式软件开发流程中的 4 个测试阶段

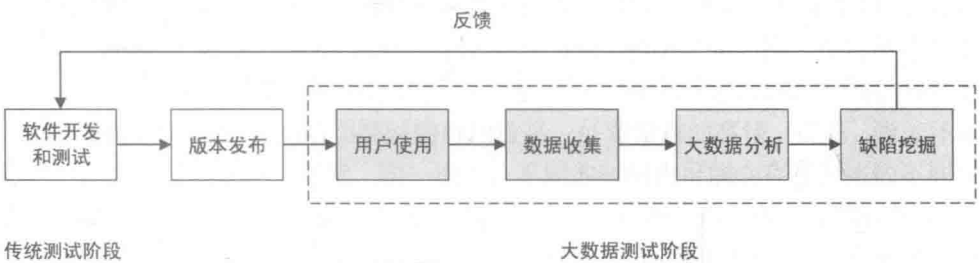


图 1.2　大数据测试的流程图

大数据测试思想的核心是通过分析海量用户的使用数据来发现传统软件测试阶段不易检测出来的软件缺陷，而不是单纯地从技术角度出发设计测试用例，探测软件缺陷。注意：在实践中，大数据测试方法并不能替代传统的软件测试方法，即使探测出了软件缺陷，仍需要软件测试人员采用传统的软件测试方法设计测试用例，进而发现软件错误的位置并修复。

1.4　目的及组织结构

为了让读者更好地理解并运用大数据技术实现软件测试的过程，本书将从工程的角度出发，通过案例来详细讲解大数据技术在软件测试中的应用，进而帮助读者建立基于大数

据分析的软件测试思维和测试方法，跳出从技术角度进行软件测试的传统思维模式。

本书的组织结构如图 1.3 所示。

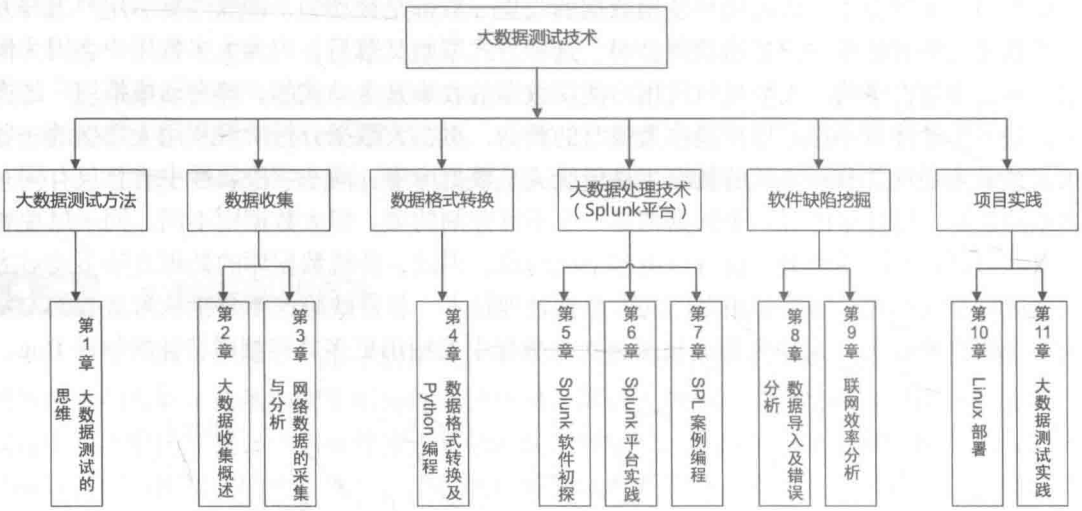


图 1.3　本书组织结构图

本书随后的章节将从数据的采集方法、数据格式转换、大数据处理平台（Splunk 平台）及其应用、软件缺陷挖掘和项目实践等 5 部分进行实例讲解。为了让本书的内容通俗易懂且具有可操作性，本书有意弱化了大数据相关方面的理论部分，突出强调了实际应用。通过实践案例，读者可掌握应用大数据技术进行软件测试的方法和过程。

扫一扫在线测

第 2 章 大数据收集概述

现在许多软件公司都非常重视收集用户的使用数据，特别是在软件崩溃之后，软件一般都会提示用户输入问题描述及发生错误的具体情况，然后发送到公司网站，然而，很多用户在软件崩溃后并不会发送这些错误信息。本书介绍了一种主动式的用户使用数据收集方式，模拟用户在不同的环境中使用手机 APP 软件，再主动收集手机 APP 软件的联网数据，并对这些数据进行分析，最终实现对手机 APP 软件的联网性能的测试。

本章将介绍如何利用手机、360 随身 Wi-Fi 和 Wireshark 软件抓取手机 APP 进行网络传输时的数据包，如何利用 Total Control 工具来实现 PC 端对手机的控制，最终实现对手机 APP 软件联网的数据收集。为了实现上述内容，本章将通过一个案例来介绍网络数据包的抓取及分析。

2.1 案例介绍

在使用手机 APP 时，用户常常会遇到一种情况，在某个地方（如便利店中）使用手机微信收发信息是正常的，但此时使用另外一款 APP 可能提示网络不可用。在一天当中，计算机连接的网络是固定的，而手机网络则随着手机持有者所在的位置不同而不同，总是在不同基站之间、不同的 Wi-Fi 之间或者是基站网络与 Wi-Fi 网络之间切换。在这种状况下，手机 APP 必须对网络切换适应性能进行必要的优化，保证在网络切换过程中手机 APP 的可用性。但探测某款手机 APP 在网络切换时的可用性是一件非常难的事情。首先，测试环境是千差万别的；其次，即使是同一环境、同一地点，不同的时间也可能会存在网络的不同稳定情况。因此，单纯地在某一个地点进行一次或多次测试难以探测手机 APP 在网络切换时的可用性，必须对大量的手机 APP 连接网络的数据进行分析，才能探测手机 APP 在网络切换时的可用性问题。

2.2 原理及方法

这里将采用 360 随身 Wi-Fi 将手机与计算机进行网络连接，随后通过 Wireshark 软件来抓取手机 APP 进行网络传输时的数据包。360 随身 Wi-Fi 利用计算机的网络来创建无线热点，只需在一台装有无线网卡的计算机上安装免费 Wi-Fi 驱动，启用后即可自动分享免费的 Wi-Fi 信号，供其他手机、笔记本电脑或移动端接入并上网。注意：在使用 360 随身 Wi-Fi 创建无线热点时，需要计算机能够连接到网络，需要安装 360 随身 Wi-Fi 驱动程序。

Wireshark（早期版本为 Ethereal）是一款网络封包分析软件，用于抓取并显示网络封包的详细信息。Wireshark 使用 WinPCAP（Windows Packet Capture）作为接口，直接与网卡进行数据报文交换。

Total Control 是一款安卓手机投屏软件。基于这款软件，可以实现计算机对手机的控制，将手机投屏到计算机，并从计算机端进行反向控制和操作手机，手机的一切功能均由计算机操作实现。

2.3 环境配置

2.3.1 配置无线网

以 Windows 10 系统为例，在桌面上右击【计算机】图标，选择【属性】命令，打开【系统】窗口，在左侧窗格中选择【设备管理器】，打开【设备管理器】窗口，选择【网络适配器】选项，找到无线网卡，如图 2.1 所示。

图 2.1 计算机中的无线网卡

在 Windows 10 操作系统中，有时需要对无线网卡的驱动进行降级，在 Windows 7 操作系统中可以跳过这步。右击【Intel(R) Wireless-N 7260】选项，选择【更新驱动程序】命令，在打开的对话框中选择【浏览我的计算机以查找驱动程序软件】选项，再在打开的对话框中选择【让我从计算机上的可用驱动程序列表中选取】选项，单击【下一步】按钮后，在打开的窗口中选择一个先前的版本进行驱动降级即可，操作过程如图 2.2 所示。

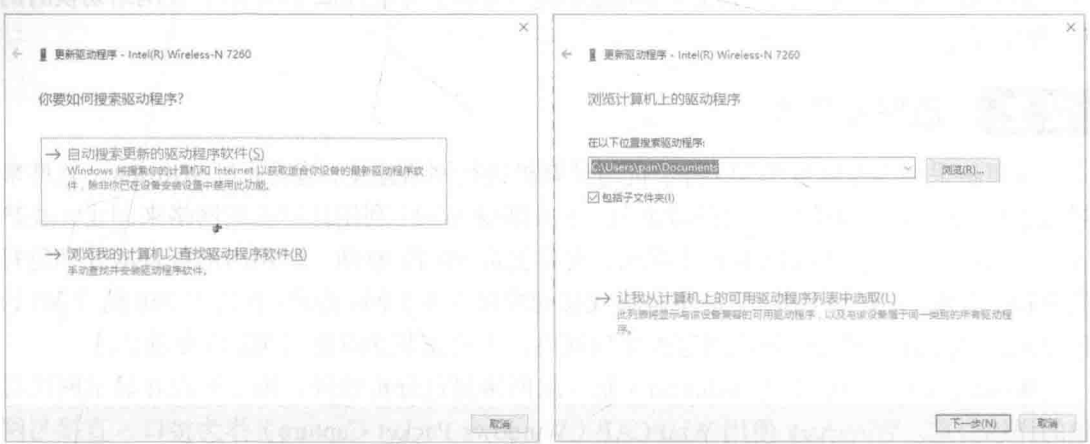


图 2.2 寻找合适的无线网卡驱动程序