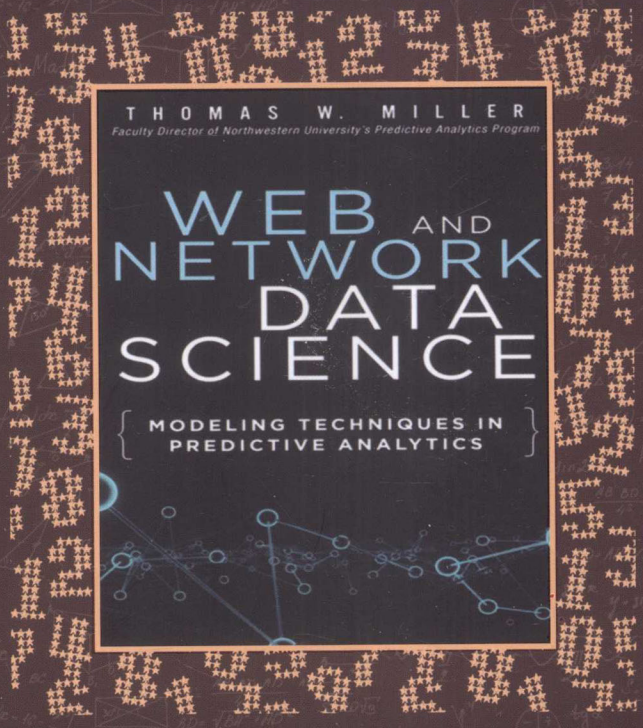
HZ BOOKS 华章IT
Pearson
数据科学与工程技术丛书

Web与网络数据科学

建模技术在预测分析中的应用

[美] 托马斯 W. 米勒（Thomas W. Miller）著
何泾沙 等译



WEB AND NETWORK DATA SCIENCE

MODELING TECHNIQUES IN PREDICTIVE ANALYTICS

机械工业出版社
China Machine Press

数据科学与工程技术丛书

WEB AND NETWORK DATA SCIENCE

MODELING TECHNIQUES IN PREDICTIVE ANALYTICS

Web与网络数据科学

建模技术在预测分析中的应用

[美] 托马斯 W. 米勒（Thomas W. Miller）著

何泾沙 等译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Web 与网络数据科学：建模技术在预测分析中的应用 /（美）托马斯 W. 米勒（Thomas W. Miller）著；何泾沙等译 . —北京：机械工业出版社，2017.1
（数据科学与工程技术丛书）
书名原文：Web and Network Data Science: Modeling Techniques in Predictive Analytics

ISBN 978-7-111-55844-6

I. W… II. ① 托… ② 何… III. 网络数据库 IV. TP311.132

中国版本图书馆 CIP 数据核字（2017）第 002923 号

本书版权登记号：图字：01-2016-2918

本书以作者在美国西北大学开设的“Web 网站分析学”课程为基础，介绍了可用性测试、网站性能、使用分析、社交媒介平台、搜索引擎优化（SEO）等方面的知识。同时，书中在涵盖实际应用与介绍社交网络分析和网络科学领域中现有的最新知识之间取得了一个良好的平衡，清楚地展示出如何将所涉及的理论知识应用于解决实际的商业问题。

本书可供计算机及相关专业学生阅读，也可供数据研究人员和分析师学习参考。

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：关　敏　　责任校对：董纪丽
印　　刷：北京瑞德印刷有限公司　　版　　次：2017 年 2 月第 1 版第 1 次印刷
开　　本：185mm × 260mm　1/16　　印　　张：16.75
书　　号：ISBN 978-7-111-55844-6　　定　　价：79.00 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

译 者 序

当今社会是一个快速发展的社会，科技发达、信息流通，人们之间的交流越来越密切，居民的生活越来越方便，大数据就是这个高科技时代的最新产物，近年来迅速成为全球IT行业中的热门词汇。大数据中所隐含的理念以及潜在的发展前景与价值已经得到越来越广泛的认可，影响着政治和经济社会中的各个方面，被认为是各类组织和机构乃至国家层面的重要战略资源，成为提高核心竞争力的有力武器，也理应得到我们每一个用户、每一个消费者的高度重视。大数据具有数据量大、种类繁杂、实时性强、蕴藏的潜在价值大等特征，公开与分享已经成为大势所趋。然而，如何鉴别数据的真伪？如何从价值密度稀疏的大数据中获取隐藏在其中的真正价值？这些疑问给人们提出了技术上的巨大挑战。Web是大数据的一个重要来源，我们每一个人通过个人电脑、手机或各类移动终端敲击的每一个字、选择的每一条词语、录制的每一段语音留言、浏览的每一个网页，都成为大数据的组成部分，进入巨大的数据海洋中，成为被提取、分析、使用的基本元素，成为形成各种商业决策的依据以及通过分析对未来可能发生的事件进行预测的基础支撑。社会信用体系中政务诚信、商务诚信、社会诚信的构建也将建立在大数据的基础之上。近年来日渐流行的社交网站在给广大用户提供即时沟通交流工具以及形成在线社区平台的同时，更是成为大数据的一个重要来源。在当今社会的线上线下，数据无处不在、持续产生，然而，众多的数据纷繁杂乱，数据之间存在的关系复杂而不明朗，我们需要去搜索、处理、分析和归纳，以挖掘出数据的深层次规律以及数据之间存在的相互关系。

大数据的价值不仅仅在于数量巨大，通过建立新的模型、提出新的方法、构建新的系统、开发新的工具对大量、动态、持续产生的结构化、半结构化和非结构化数据进行分类、融合、分析与挖掘，以获得具有实际应用价值以及能够预测未来事件与行为的结果，这才是大数据的真正价值所在。虽然数据的迅速膨胀将决定企业、机构的未来发展，然而很多企业和机构并没有意识到数据爆炸性增长所带来的机遇以及潜在的隐患。但是随着时间的推移，随着大数据分析技术的进一步成熟与完善，大数据将得到越来越多的应用，实现越来越大的价值，人们也将越来越多地意识到大数据在企业和机构运作中所起的重要作用。在未来的商业、经济及其他领域中，关键的决策将越来越多地基于大数据与数据分析而做出，而越来越少地基于经验和直觉。

大数据将在观念上给我们带来一些颠覆性的转变。首先，我们面对的将是全部数据，而不再是随机抽取的样本；其次，大数据提供给我们的是混杂性，而不再是精确性；再次，

大数据之间存在的是相关关系，而不再只是因果关系。因此，对大数据的分析需要我们提出和构建新的模型、方法和工具。本书就是为了满足这些新的需求和新的要求而撰写的，是将数据科学与网络科学相结合形成的“Web 与网络数据科学”。本书不但包含了大数据分析与应用所需的理论知识与建模技术，还提供了大量的应用实例，并通过提供建模技术方面的资料和参考指南对研究人员及分析师的工作提供进一步的帮助，同时面向实际应用向编程人员展示如何使用目前在数据分析与应用领域中得到广泛应用的 Python 和 R 语言编写能够正确运行并解决实际商业问题的计算机软件，还提供了大量的 Python 和 R 语言代码实例。全书涵盖了 Web 与网络数据科学领域中的若干主要问题，如网站设计与用户行为、网络路径与通信、社区与影响、个体与群体行为、信息与网络等，具体分 12 章，在对开展 Web 与网络数据科学方面的研究所需的相关技术进行概述后，分别对 Web 在线消息传递技术、Web 爬行与抓取技术、Web 链接测试以及体验与外观改进技术、在线竞争性情报搜集与分析技术、网络可视化技术、社区发现与分析技术、情感度量技术、基于文本的共同主题发现技术、推荐技术、网络博弈行为的建模技术进行了深入浅出的介绍，最后对未来 Web 的发展进行了展望。此外，本书使用较大篇幅，以附录的方式对目前数据建模与分析中常用的技术进行了简要介绍，包括数据库与数据准备、数据统计学、回归与分类、机器学习、数据可视化以及文本分析学，对开展在线研究的流程与方法进行了系统的归纳，最后通过提供若干实用案例为本书中介绍的理论知识和应用技术画上了一个完美的句号。本书还向读者提供了在应用预测分析学的建模技术中常用的代码与共享程序、常用术语以及丰富的参考文献，为读者进一步学习提供专业帮助与技术指南。因此，本书对于从事基于 Web 的数据搜集、分析和应用的技术人员以及在相关领域中从事科学探索和技术研发的科研人员具有较重要的参考价值。

本书由三峡大学“楚天学者计划”主讲教授、北京市特聘教授何泾沙博士负责翻译，三峡大学贺鹏教授、中国科学院软件研究所朱娜斐博士协助了全程的翻译工作，中国航天科技集团公司第九研究院第十三研究所张玉强博士、清华大学徐晶博士、中国科学院信息工程研究所徐菲博士、北京工业大学博士生赵斌、研究生朱星烨、方静、刘畅、黄辉祥参与了部分章节的翻译工作。何泾沙博士对全书进行了最终统稿及全文校验。由于译者的水平有限，再加上时间方面的限制，译文中难免存在不够准确之处，敬请广大读者批评指正，译者在此深表谢意。

何泾沙

2016 年 12 月于北京

前　言

“斯考特，把我弹射出去。”

Captain Kirk（William Shatner 饰）
电影《星际旅行 4：抢救未来》（1986 年）

Web 是一个由众多网页相连接而形成的网络，是一个通信媒介，是一个覆盖全球的信息来源。人们花费大量的时间在 Web 上进行搜索，获取有用的数据与信息，并对它们进行分析。有效使用 Web 给人们的生活带来了很多的便利。本书将告诉你以上这一切是如何实现的。

本书是根据我在西北大学（Northwestern University）讲授的一门课程的内容撰写而成的。此课程从介绍 Web 网站分析学入手，主要关注在 Web 搜索中使用数据的统计与性能。之后，我又在此课程中增加了来自网络科学和社交媒体的概念。在讲授此课程两年后，我认识到从 Web 上收集信息可以成为一个独立的话题，有太多关于 Web 与网络数据科学方面的知识可以学习。本书就像我讲授的课程那样，是关于以上这些知识的指南。

Web 与网络数据科学是数据科学和网络科学相结合而形成的，关注的是将 Web 看成一个提供信息的来源。因而，最好的学习方法就是通过实例进行讲解。因此，本书中包含大量的实例，通过提供建模技术方面的资料和参考指南给研究人员与分析师提供帮助。我们也会向编程人员展示如何基于基础代码编写能够正确运行并用于解决真实商业问题的软件。

我们想要做的事情都会通过所编写的代码体现出来。本书中包含的这些代码将作为参考资料提供给每一位读者，当然会有部分读者对这些代码进行进一步调试。为了鼓励学生学习，每一段程序代码都包含详细的注释以及如何进一步分析的建议。所有的数据集以及计算机程序代码都可以直接从本书的网站 http://www.ftpress.com/miller/ 下载。

Python 这个名字来源于 Monty Python。大家会看到有些软件包的名称比较奇特，如 Twisted 或 Scrapy。R 语言拥有自己的 lubridate 与 zoo 软件开发包。好的结果来源于辛勤工作并热爱工作的人们。那些追求快乐而不是名利的人们为开源软件做出了贡献，而我很高兴自己能够成为开源软件 Python 和 R 语言社区中的一员。那就让我们一起开始这段快乐的旅程吧！

对于 Web 和网络中存在的问题，使用 Python 可以有效便捷地解决某些问题，而使用

R 语言可以有效便捷地解决其他一些问题。常常还会出现两种语言都适用的情况，这时就需要进行权衡。总体来说，Python 和 R 语言能够用于对 Web 及网络数据进行有效的收集与分析。

在本书中，我们还会提到编程时会使用到的很多工具。对网站的正常运行负有责任的 Web 专业技术人员还会使用很多其他语言和技术，如 JavaScript、Apache、.Net Web 服务，以及数据库系统。本书的讨论将会涉及这些技术，但不会提供任何编程代码。

本书中大多数数据来源于公共域数据源。用于支持案例的数据来源于加利福尼亚大学尔湾分校的机器学习信息库（Machine Learning Repository）和斯坦福大学的大型网络数据集（Large Network Dataset Collection）。所获取的影视方面的数据得益于互联网影视数据库（Internet Movie Database）所给予的使用许可。IMDb 影视评价数据由斯坦福大学的 Andrew L. Mass 及同事整理完成。安然（Enron）案例数据由卡耐基 – 梅隆大学的 William W. Cohen 维护。Quake Talk（地震谈话）案例数据由 Maksim Tsvetovat 维护。我们对以上这些学者为我们的研究提供了丰富的数据表示深切的感谢。

很多人对我这些年来的知识积累都产生过重大的影响。他们中有出色的思考者，有善良的同仁，还有我会永远感激的老师以及导师。不幸的是，尤西纽斯学院（Ursinus College）哲学系的 Gerald Hahn Hinkle 和语言系的 Allan Lake Rice 以及明尼苏达大学（University of Minnesota）哲学系的 Herbert Feigl 已经永远离开了我们。在此，我还要感谢明尼苏达大学心理测验学系的 David J. Weiss 以及曾经在俄勒冈大学（University of Oregon）经济系任教的 Kelly Eakin。好的老师（没错，他们都是伟大的园丁）终身都将得到人们的尊重。

感谢 Stan Narusiewcz 给了我职业生涯中的第一份工作，那是一个网络工程师的岗位。感谢 Tom Obinger 指导我成为一个成功的计算机系统和网络销售人员。还有 Bill JoBush 和 Brian Hill，在我作为信息系统专业人员整个职业生涯的各个阶段，他们曾经是我的直接上司或同事。

感谢 Michael L. Rothschild、Neal M. Ford、Peter R. Dickson 和 Janet Christopher 在威斯康星大学麦迪逊分校（University of Wisconsin–Madison）伴我一起度过几年美好的时光并给予我无私的帮助。特别感谢 A. C. Nielsen Center for Marketing Research 的学生和顾问委员会的专家以及 Jeff Walkowski 和 Neli Esipova，后两位在我组织在线调查与专题讨论小组期间曾经同我一起工作，我们所使用的方法那时才开始在重要的研究中得到应用。

我很有幸参与了西北大学成人教育学院开展的研究生远程教育的课程教学活动。感谢 Glen Fogerty 给我提供了讲授课程的机会，并让我负责西北大学预测分析学项目。感谢所有参与这个很有特色的研究生项目的同事和管理人员。最后，感谢帮助过我的众多学生们和老师们，你们令我受益匪浅。

ToutBay 是数据科学领域中一个快速成长的公司。与公司的共同创始人 Greg Blence 一样，我对公司的未来发展抱有很大的信心。感谢 Greg 让我有这样一个参与创业以及面

对商业活动中的现实而能够更加脚踏实地的机会。学术以及数据科学模型毕竟有其局限性，为了能够真正产生影响，我们必须实现我们的想法和模型，并且与他人进行共享。

我的家在加利福尼亚州，道奇体育馆（Dodger Stadium）以北四英里[⊖]，但是我在位于伊利诺伊州埃文斯顿市（Evanston, Illinois）的西北大学任教，同时在位于佛罗里达州坦帕市（Tampa, Florida）的一个名叫 ToutBay 的数据科学公司指导产品研发。这样的工作和生活方式充分体现出了互联网带给我们的巨大便利。

TEXnology 公司的 Amy Hendrickson 使本书的编排、文字、图表看上去都是那么出色和完美，这是开源软件的又一个成功实例。感谢 Donald Knuth 以及 TEX/LATEX 整个社区对这个出色的系统在编排和出版方面做出的贡献。

本书中包含的内容主要源于在西北大学讲授的 Web 与网络数据科学这门课程。参与课程学习的学生提出了很多想法和启示。Lorena Martin 对本书进行了评阅，提供了许多宝贵意见。Candice Bradley 不但评阅了本书，还是本书的文字编辑。我对他们给予的帮助和鼓励表示衷心感谢。最后还要感谢我的编辑 Jeanne Glasser Levine 以及本书的出版商 Pearson/FT Press，是他们使本书的成功出版成为可能。在此特别声明，我个人对所有写作方面的事宜、存在的错误与问题以及不足负全部责任。

我的好朋友 Brittney 和她的女儿 Janiya 总是抽空陪伴我。我的儿子 Daniel 总能与我同甘共苦，是我一辈子的朋友。我对于他们给予的信任致以崇高的敬意。

Thomas W. Miller

美国加利福尼亚州格伦代尔市

⊖ 约 6.4 公里。——编辑注

目　录

第 1 章
相关技术概述

你为什么不抽时间来看我？

Lady Lou（Mae West 饰）
电影《侬本多情》（1933 年）

我的职业生涯起始于明尼苏达州罗斯维尔市的一名网络工程师。从明尼苏达大学统计专业研究生毕业后，我具备了很好的数学和建模方面的基础，但是缺少商业方面的知识。我很快就意识到要在职业发展上取得成功就需要掌握更多管理方面的知识。

20 世纪 70 年代末是一个通过拨号或专线上网的时代，异步通信、半同步通信及同步通信是当时的主流。我们将网络协议通过轮询方式和信息比特位来表达，并标注出每一条通信线路能够承载的传输速率。排队理论（Queuing Theory）及离散事件模拟是开展以上分析的理论基础。

银行里的柜台职员会提交一个请求，然后按下终端上的回车键。终端连接到一个控制器，控制器再连接到远端的一个汇聚处理器。汇聚处理器再通过专线连接到一个前端处理器，从而建立起一个连接到大型计算机的通道。以上就是那个年代网络中的节点和连接。排队理论用于估算银行的柜台职员需要等待多久才能得到大型计算机的响应。

40 年飞快地过去了。我们已经远离了拨号和专线的时代。今天通信协议的基础是数据包的交换以及移动通信。使用网络的用户已经遍及各个角落，而不再只局限于银行、公司和研究机构。绝大多数大型计算机也被小型计算机群所取代。我们的口袋里装着最小的计算机。如果愿意，我们还可以将计算机穿戴在身上。然而，当我们向远端的系统提交请求时，我们仍然需要等待响应，不同的是我们现在可以在任何地方等待，同时可以做些其他的事情。

随着计算机硬件方面的差别逐渐消失以及软件走向开源化，现有的科技公司都在寻找商业智能与数据科学方面的商机。IBM 从一个以生产硬件为主的公司转型成为一个软件开发商，然后又转型成为一家咨询公司。HP（惠普公司）已经一分为二，一部分专营硬件，另一部分则以提供商业服务和工具为主要经营范围。同时，苹果（Apple）与亚马逊（Amazon）和谷歌（Google）在多媒体发布领域中展开激烈竞争，并对三星公司（Samsung）违反软件著作权提

起法律诉讼。

今天主要的商业竞争都涉及信息以及信息的在线发布。知识产权、专业知识、竞争智能、专长及艺术都给在线市场增加了很多新的价值，不考虑以上因素的话，在线获取信息基本上都可以免费实现。

人们很难抵挡Web的诱惑，因为它为用户提供了拥有无穷无尽信息的空间，还可以提供能够到达所有这些信息的连接。Web是一个巨大的数据仓库，是一条通向知识的道路，更是一个开发新知识的研究媒介。

Web与网络数据科学由多项科学技术以及建模技术所组成，其中某些技术已经相当成熟，还有些新技术仍然处在发展与完善当中，这些技术能够帮助我们了解生活中的Web及网络。Web也有多项技术得到应用，Alexa Internet公司（2014）、W3Techs公司（2014）等企业专门对各项技术的市场占有率进行跟踪。

为了更有效地开展Web与网络数据科学方面的研究，需要具备相应的技术背景，至少需要对Python、R语言和JavaScript（Java脚本语言）有一定的了解。Python是一款进行数据预处理的必选工具（有时也称为数据改造）。R语言是一款专门用于数据建模和可视化的工具。JavaScript是Web客户端开发中使用的一种语言，主流的Web浏览器都可以支持。在解决Web和网络方面的问题中，以下方面的知识和技能也非常有用：HTML5、CSS3、
2 XPath、各种文本和图像文件格式、Java、Linux、Apache、.Net Web服务、数据库系统以及服务器端开发中使用的语言，如Perl、PHP。虽然需要具备上面提到的相关专业知识，但是一本书能够涵盖的内容毕竟有限。因此，我们将会在本书的最后增加一个有关术语表的附录。

自Brendan Eich于1995年在网景公司（Netscape）工作期间开发出JavaScript以来，作为一种开发语言，JavaScript迅速成为Web客户端开发所使用的语言，是一款基于浏览器对用户交互进行管理的引擎。JavaScript在客户端占有垄断地位，88%的网站使用该技术，另外11.8%的网站仍然使用静态的纯HTML技术，无法为客户端提供任何编程能力（W3Techs 2014）。

2008年，Crockford总结了JavaScript的优缺点，而更多的人则告诉我们如何在实际系统中去使用它（Stefanov 2010; Flanagan 2011; Resig and BearBibeault 2013）。近年来，随着Node.js的出现，JavaScript在服务器端也开始得到应用（Hughes-Croucher and Wilson 2012; Wanderschneider 2013; Cantelon, Harter, Holowaychuk, and Rajlich 2014）。有些人大力推动端到端的JavaScript应用程序，即在客户端和服务器端都使用JavaScript程序以及文件数据库（Mikowski and Powell 2014）。JavaScript Object Notation（JSON）提供了一种数据交换格式，可读性比XML更好，并可以很容易地与任何MongoDB文件数据库进行集成（Chodorow 2013; Copeland 2013; Hoberman 2014）。如果具有作为一个建模与分析语言所需要的功能，JavaScript肯定也能在Web领域发挥主导作用。不幸的是，JavaScript并没有做到。

今天的数据科学领域吸引了能够熟练使用R语言的统计专家和能够熟练使用Python的信息技术专家。这两类研究人员还有很多需要相互学习之处。对于想要真正开展实际工作的

数据科学家来说，熟悉以上两种语言会使他们具备更多的优势。

R 语言由 Ross Ihaka 和 Robert Gentleman 在 1993 年设计并推出，是一款可扩展、面向对象的开源脚本语言，用于通过编程对数据进行处理。此语言在数据统计界得到广泛应用，其语法、数据结构及编程方法与它的前身 S 语言和 S+ 语言相似。此语言的贡献者为广泛应用提供了超过 5 000 个软件开发包，主要提供传统的数据统计、机器学习及数据可视化方面的功能。R 语言目前是数据科学界使用最广泛的编程语言，但它并不是一个通用目的的编程语言。

3

Guido van Rossum（Monty Python 的一个追随者）在 1994 年设计并发布了 Python 1.0 版本。这个通用目的编程语言在随后的年代里慢慢流行起来。很多系统编程人员都从先前使用 Perl 改为使用 Python，Python 尤其获得从事数学和自然科学的研究人员的青睐。很多高等院校将 Python 作为介绍面向对象程序设计语言基本概念的一种手段。一个非常活跃的开源社区贡献了超过 15 000 个 Python 软件开发包。

Python 时常会被誉为一种“胶合语言”，它为科学编程与科研提供了一个非常丰富的开源环境。对于需要占用大量计算机资源的应用程序来说，Python 提供了调用通过正确编译的 C、C++ 和 Fortran 子程序的功能。我们也可以使用 Cython 将 Python 代码转换成优化后的 C 语言代码。对于没有在 Python 中实现的建模技术或图像来说，我们可以在 Python 程序中调用 R 语言程序。

有些问题可以使用 Python 很容易地解决，其他一些问题则可以使用 R 语言很容易地解决。Python 作为一个通用目的编程语言给我们提供了很多便利，并能通过使用 R 语言编程软件包得到传统的数据统计、时间序列分析、多元方法、统计图表制作以及丢失数据处理方面的功能。因此，本书将包含 Python 和 R 语言的代码实例，是一部在 Web 与网络数据科学领域同时使用这两种语言的指导书籍。

浏览器的使用在发展过程中经历了很大的变化，随着 Google（谷歌）Chrome 浏览器市场份额的增加，微软（Microsoft）Internet Explorer（IE）的使用呈现下降趋势。表 1-1 和图 1-1 展现的是 2008 年 10 月～2014 年 10 月全球浏览器的使用情况，图表中的数据来源于 StarCounter (2014)。熟练使用浏览器以及由浏览器提供的查阅文本元素和网页结构的工具非常有帮助。

表 1-1　全球 Web 浏览器占有率统计（2008～2014）

年	IE	Chrome	Firefox	Safari	其　他
2008	67.68	1.02	25.54	2.91	2.85
2009	57.96	4.17	31.82	3.47	2.58
2010	49.21	12.39	31.24	4.56	2.60
2011	40.18	25.00	26.39	5.93	2.50
2012	32.08	34.77	22.32	7.81	3.02
2013	28.96	40.44	18.11	8.54	3.95
2014	19.25	47.57	17.00	10.95	5.23

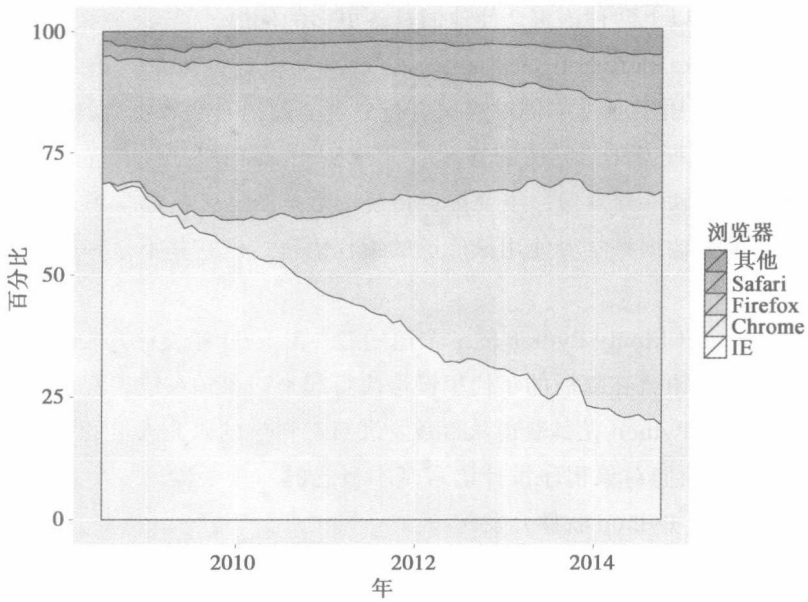


图 1-1　全球 Web 浏览器使用情况（2008.7～2014.10）

“大数据”带来的挑战不仅仅是数据的数量，而是我们对数据产生的来源缺乏足够的了解，特别是 Web 以及社交媒体。数据存在于 Web 的各个角落。我们需要能够找到有关数据并获取这些数据的高效方法。

应用程序编程接口（API）是一个从 Web 获取数据的方法。Russell（2014）对社交媒体 API 做了一个整体回顾。不幸的是，调用 API 对语法、参数及授权码都有要求，数据提供者可以随意进行修改。我们采取的方法有所不同，关注重点是从 Web 自动获取数据的通用技术。

图 1-2 对在线研究过程进行了总结。数据采样、收集及准备会耗费大量的时间，使二次研究超过初步研究而处于主导地位。在线二次研究以 Web 中已经存在的数据为基础。我们将在第 3 章介绍二次研究方法，这些方法将在随后的章节中应用。在线初步研究得到了 Web 的支持，我们将会在附录 B 中对相关的方法进行介绍。

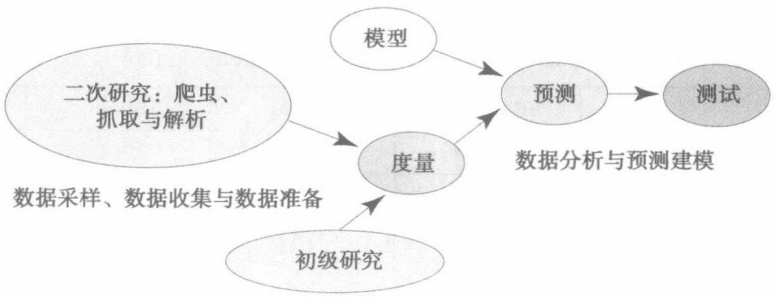


图 1-2　Web 与网络数据科学：在线研究过程

Web 与网络数据科学的范围很大，有众多问题亟待解决，主要问题如下。

- **网站设计与用户行为**。正如很多人理解的那样，Web 分析学涉及对某个特定网站的用户数据进行收集、存储与分析。这就带来了很多需要解决的问题。如何设计和实现网站（达到易用、可视性好、市场宣传、高效搜索、将访问转化为销售等目的）？如何从 Web 中高效获得信息？如何将半结构化和非结构化的文本转换为可以用于分析和建模的数据？哪些方法适合对网站及社交媒体进行度量？谁在使用网站，他们如何使用网站？网站提供的服务是否满足用户的需要？如何将一个网站与其他网站进行比较？ 5～6
- **网络路径与通信**。Web 与网络数据科学远远超出网站分析学的范畴。我们对网站的审视或评价都是相对于 Web 中的其他网站。我们围绕着网络进行思考，即提供信息的节点之间相互连接、用户之间相互通信。那么，两个节点之间的最短路径、最快路径或成本最低路径在哪里？通过网络传递消息最快的方法是什么？在完成一个项目的关键路径上都有哪些活动？从服务器得到响应需要多长时间？
- **社区与影响**。电子社交网络从社交媒体可见一斑。在这个领域中，存在很多关于社交网络分析方面的问题。在某个社区中是否存在可识别的用户群体？哪些用户是群体中关键或最重要的参与者？拥有声望、影响力或权利的用户在哪里？哪些用户在成长为群体领导者的过程中处于最有利的位置？
- **个体与群体行为**。作为一个研究数据的科研人员，要求做到的不仅仅是描述现象而已，而是要对将来的行为或性能进行预测。因此，有更多的问题需要我们解决。已知一个用户与其他购买者或非购买者之间的关系，那么这个用户也会购买吗？已知一个用户与其他投票人的关系，这个用户也会投票给某位候选人吗？已知个人动机，一个群体接下来会做什么呢？已知网络过去的发展，我们预测网络将来会怎样发展呢？
- **信息与网络**。作为信息的来源，Web 是无可比拟的。在线信息属性方面的问题由而产生。哪些网站是获取某方面信息的最好来源？谁是提供信息的最可靠来源？如何归纳出领域知识的特征？如何使用 Web 获得具有竞争力的智能？如何像使用数据库一样将基于 Web 的信息用来回答问题（专业问题及通用问题）？

本书的目的就是对 Web 与网络数据科学领域进行概括性介绍。我们将介绍为了回答问题而需要使用的度量与建模技术，并提供进一步学习的参考资料。介绍的技术从基础到超前，然而它们对于数据科学研究来说都非常重要。 7

某些人认为所谓的数据科学其实就是一些新的统计方法。在一个由数据主宰的世界里，数据科学同时似乎开始提供新的商机以及新的信息技术。在需要解决 Web 和网络问题时，后面这一点显得尤为重要。由 Web 产生和传输的无穷无尽的数据将会使这些研究持续相当长一段时间。

作为本书的首个软件程序，代码 1-1 是一段展示 Web 浏览器使用统计数据的 Python 程序。代码 1-2 是使用 R 语言编写的与之相对应的程序，其中用到了由 Wickham 和 Chang（2014）编写的制图软件。 8

代码 1-1 浏览器使用情况分析（Python）

```
# Analysis of Browser Usage (Python)

# prepare for Python version 3x features and functions
from __future__ import division, print_function

# import packages for data analysis
import pandas as pd  # data structures for time series analysis
import datetime  # date manipulation
import matplotlib.pyplot as plt

# browser usage data from StatCounter Global Stats
# retrieved from the World Wide Web, October 21, 2014:
# \url{http://gs.statcounter.com/#browser-ww-monthly-200807-201410
# read in comma-delimited text file
browser_usage = pd.read_csv('browser_usage_2008_2014.csv')
# examine the data frame object
print(browser_usage.shape)
print(browser_usage.head())

# identify date fields as dates with apply and lambda function
browser_usage['Date'] = \
    browser_usage['Date']\
    .apply(lambda d: datetime.datetime.strptime(str(d), '%Y-%m'))
# define Other category
browser_usage['Other'] = 100 -\
    browser_usage['IE'] - browser_usage['Chrome'] -\
    browser_usage['Firefox'] - browser_usage['Safari']

# examine selected columns of the data frame object
selected_browser_usage = pd.DataFrame(browser_usage,\
    columns = ['Date', 'IE', 'Chrome', 'Firefox', 'Safari', 'Other'])
print(selected_browser_usage.shape)
print(selected_browser_usage.head())

# create multiple time series plot
selected_browser_usage.plot(subplots = True,  \
     sharex = True, sharey = True, style = 'k-')
plt.legend(loc = 'best')
plt.xlabel('')
plt.savefig('fig_browser_mts_Python.pdf',
    bbox_inches = 'tight', dpi=None, facecolor='w', edgecolor='b',
    orientation='portrait', papertype=None, format=None,
    transparent=True, pad_inches=0.25, frameon=None)

# Suggestions for the student:
# Explore alternative visualizations of these data.
# Try the Python package ggplot to reproduce R graphics.
# Explore time series for other software and systems.
```


代码 1-2 浏览器使用情况分析（R 语言）

```
# Analysis of Browser Usage (R)

# begin by installing necessary package ggplot2

# load package into the workspace for this program
library(ggplot2)  # grammar of graphics plotting

# browser usage data from StatCounter Global Stats
# retrieved from the World Wide Web, October 21, 2014:
# \url{http://gs.statcounter.com/#browser-ww-monthly-200807-201410
```

```
# read in comma-delimited text file
browser_usage <- read.csv("browser_usage_2008_2014.csv")
# examine the data frame object
print(str(browser_usage))
# define Other category
browser_usage$Other <- 100 -
    browser_usage$IE - browser_usage$Chrome -
    browser_usage$Firefox - browser_usage$Safari

# define time series data objects
IE_ts <- ts(browser_usage$IE, start = c(2008, 7), frequency = 12)
Chrome_ts <- ts(browser_usage$Chrome, start = c(2008, 7), frequency = 12)
Firefox_ts <- ts(browser_usage$Firefox, start = c(2008, 7), frequency = 12)
Safari_ts <- ts(browser_usage$Safari, start = c(2008, 7), frequency = 12)
Other_ts <- ts(browser_usage$Other, start = c(2008, 7), frequency = 12)

# create a multiple time series object
browser_mts <- cbind(IE_ts, Chrome_ts, Firefox_ts, Safari_ts, Other_ts)
dimnames(browser_mts)[[2]] <- c("IE", "Chrome", "Firefox", "Safari", "Other")
# plot multiple time series object using standard R graphics
pdf(file="fig_browser_mts_R.pdf",width = 11,height = 8.5)
ts.plot(browser_mts, ylab = "Percent Usage", main="",
    plot.type = "single", col = 1:5)
legend("topright", colnames(browser_mts), col = 1:5,
    lty = 1, cex = 1)
dev.off()

# define Year as numeric with fractional values for months
browser_usage$Year <- as.numeric(time(IE_ts))

# build data frame for plotting a stacked area graph
Browser <- rep("IE", length = nrow(browser_usage))
Percent <- browser_usage$IE
Year <- browser_usage$Year
plotting_data_frame <- data.frame(Browser, Percent, Year)

Browser <- rep("Chrome", length = nrow(browser_usage))
Percent <- browser_usage$Chrome
Year <- browser_usage$Year
plotting_data_frame <- rbind(plotting_data_frame,
    data.frame(Browser, Percent, Year))
Browser <- rep("Firefox", length = nrow(browser_usage))
Percent <- browser_usage$Firefox
Year <- browser_usage$Year
plotting_data_frame <- rbind(plotting_data_frame,
    data.frame(Browser, Percent, Year))

Browser <- rep("Safari", length = nrow(browser_usage))
Percent <- browser_usage$Safari
Year <- browser_usage$Year
plotting_data_frame <- rbind(plotting_data_frame,
    data.frame(Browser, Percent, Year))

Browser <- rep("Other", length = nrow(browser_usage))
Percent <- browser_usage$Other
Year <- browser_usage$Year
plotting_data_frame <- rbind(plotting_data_frame,
    data.frame(Browser, Percent, Year))

# create ggplot plotting object and plot to external file
pdf(file = "fig_browser_usage_stacked_area_R.pdf", width = 11, height = 8.5)
area_plot <- ggplot(data = plotting_data_frame,
    aes(x = Year, y = Percent, fill = Browser)) +
    geom_area(colour = "black", size = 1, alpha = 0.4) +
```

10

```
        scale_fill_brewer(palette = "Blues",
            breaks = rev(levels(plotting_data_frame$Browser))) +
        theme(legend.text = element_text(size = 15))  +
        theme(legend.title = element_text(size = 15)) +
        theme(axis.title = element_text(size = 15))
    print(area_plot)
    dev.off()
```