

硕 士 研 究 生 读 书 报 告



题目 大数据时代

作者姓名 张明政

作者学号 21651020

指导教师 李启雷

学科专业 移动互联网与游戏开发技术

所在学院 软件学院

提交日期 2017年 1月

The Big Data Era

A Dissertation Submitted to

Zhejiang University

in partial fulfillment of the requirements for

the degree of

Master of Engineering

Major Subject: Software Engineering

Advisor: Li Qilei

By

Zhang Mingzheng

Zhejiang University, P.R. China

2017

摘要

本文重点探讨了大数据在现在这个时代的重要性和粗略的介绍大数据。大数据指的是所涉及的数据量规模巨大到无法通过人工或者计算机，在合理的时间内达到截取、管理、处理、并整理成为人类所能解读的形式的信息。大数据的核心就是预测。从方方面面对大数据进行预测。随着大数据被越来越多的提及，有些人惊呼大数据时代已经到来了。大数据时代的来临带来无数的机遇，但是与此同时个人或机构的隐私权也极有可能受到冲击，大数据包含各种个人信息数据，现有的隐私保护法律或政策无力解决这些新出现的问题。

**关键词**：大数据， 预测

Abstract

The paper discuss the important of big data in this era and introduce it roughly. Big data is that enormous data can’t be calculate by people and computer in time to get, conduct, deal with and manage it. The main function of big data is prediction. We predict all kinds of things with big data. Big data is more and more popular and then some people think that it’s a big data era. The big data era has many chances, but the privacy is also impacted. Big data contains all kinds of personal data, the law can’t protect it now.

**Keywords：**big data, predict

1引言

2012年后，大数据被越来越多地提及，它已经上过《纽约时报》、《华尔街日报》的专栏封面，进入美国白宫网的新闻，并频繁出现在互联网、移动互联网、电商、电信、金融等领域的大佬们的话语中，不管是阿里巴巴的前CEO马云，还是腾讯的董事长马化腾，都在谈论着大数据。大数据指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

亚马逊、沃尔玛、雅虎、IBM、苹果、谷歌这些国际巨头更是在很早的时候，就开始在大数据方面进行研究和布局。我也对大数据比较感兴趣，于是就打算写大数据方向的读书报告。我花了一个星期的时间拜读了维克托·迈尔-舍恩伯格的《大数据时代》，受益匪浅。在[维克托·迈尔-舍恩伯格](http://baike.baidu.com/view/9716309.htm)及肯尼斯·库克耶编写的《[大数据时代](http://baike.baidu.com/subview/9424571/15364102.htm)》中大数据指不用随机分析法（[抽样调查](http://baike.baidu.com/view/68376.htm)）这样捷径，而采用所有数据进行分析处理。大数据的5V特点（IBM提出）：Volume（大量）、Velocity（高速）、Variety（多样）、Value（低价值密度）、Veracity（真实性）。大数据就是[互联网](http://baike.baidu.com/view/6825.htm)发展到现今阶段的一种[表象](http://baike.baidu.com/view/39081.htm)或[特征](http://baike.baidu.com/view/1069886.htm)而已，没有必要神话它或对它保持敬畏之心，在以[云计算](http://baike.baidu.com/view/1316082.htm)为代表的技术创新大幕的衬托下，这些原本看起来很难收集和使用的数据开始容易被利用起来了，通过各行各业的不断创新，大数据会逐步为人类创造更多的价值。

其次，想要系统的[认知](http://baike.baidu.com/view/69807.htm)大数据，必须要全面而细致的分解它，我着手从三个层面来展开：

第一层面是[理论](http://baike.baidu.com/view/96020.htm)，理论是认知的必经途径，也是被广泛认同和传播的基线。在这里从大数据的特征定义理解行业对大数据的整体描绘和定性；从对大数据价值的探讨来深入解析大数据的珍贵所在；洞悉大数据的发展趋势；从大数据[隐私](http://baike.baidu.com/view/350273.htm)这个特别而重要的视角审视人和数据之间的长久博弈。

第二层面是[技术](http://baike.baidu.com/view/45517.htm)，技术是大数据价值体现的手段和前进的基石。在这里分别从云计算、分布式处理技术、存储技术和感知技术的发展来说明大数据从采集、处理、存储到形成结果的整个过程。

第三层面是[实践](http://baike.baidu.com/view/51642.htm)，实践是大数据的最终价值体现。在这里分别从互联网的大数据，政府的大数据，企业的大数据和个人的大数据四个方面来描绘大数据已经展现的美好景象及即将实现的蓝图。

本读书报告分为三部分，第一部分是全书概述，主要包括作者简介、内容概要，让大家对《大数据时代》一书有一初步的了解；第二部分是感悟思考，主要包括大数据在开放平台、社交网络、电子商务领域的应用分析；第三部分是过程体会，主要是对读书过程及方法的一种分享，希望与大家有所交流互益。

**2 全书概述**

本书作者维克托·迈尔-舍恩伯格被誉为“大数据时代的预言家”，现任牛津大学网络学院互联网研究所治理与监管专业教授，曾任哈佛大学肯尼迪学院信息监管科研项目负责人，新加坡国立大学信息政策研究中心主任，他的咨询客户包括微软、IBM和惠普等全球顶级企业。《经济学人》说，在大数据领域，他是最受人尊敬的权威发言人之一；《科学》说，若要发起一场关于这个问题的深入讨论，没有比他更好的发起者了。他是欧盟互联网官方政策背后的重要制定者和参与者；他是最早洞见大数据时代发展趋势的数据科学家之一。

而本书的译者，现年30岁的周涛则是电子科技大学互联网科学中心主任、教师、博士生导师，发表了SCI论文150余篇，领域一流期刊90余篇。

整本书分为三个部分，分别是大数据时代的思维变革、商业变革和管理变革。

大数据时代的思维变革分为“不是随机样本，而是全体数据”、“不是精确性，而是混杂性”、“不是因果关系，而是相关关系”三部分，分别讲述了大数据时代，数据在数量、性质和关系方面的变革。当数据处理技术得到了较大的发展之后，数据分析就采取了全数据模式，样本=总体，比如淘宝进行用户行为分析时会就一个用户在淘宝里的所有行为进行技术分析，而谷歌对流感的预测则是对所有相关词汇搜索行为进行技术分析。书里说，只有5%的数据是结构化且能适用于数据库的，如果不接收混乱，剩下的95%的非结构化数据都无法被利用。用户在互联网上的行为，其实就是一种数据，比如鼠标停留位置、搜索输入语言、项目点击次数等，但是这些数据其实都不是标准化、结构化的，用户搜索板蓝根、感冒灵、吃什么水果治感冒等都说明用户可能感冒了，但是其输入的搜索文字确实不一样的，如果要精准地按照感冒两个字来判断用户是否感冒，其数据分析结果和实际结果应该会存在比较大的差距。淘宝数据有一个视频内容是各个省份的人购物情况的分析统计，比如在见不到大海的新疆，购买比基尼却是最多的，那么淘宝并不需要知道新疆人为什么喜欢买比基尼，是因为气候还是生活习惯并不重要，重要的是淘宝要多为新疆用户推荐比基尼，可以获得更高的转化率。（推荐视频《淘宝数据盛典——你所不知道的城市秘密》

大数据时代的商业变革分为内容数据化、数据价值实现、数据公司类型三部分，随着信息技术的变革，我们将文字、方位、沟通甚至是世间万物都变成数据，然后通过数据的再利用、重组、扩展等实现价值，创造大数据公司。同时，根据所提供价值的数据本身、技能和思维三种不同来源，分别出现了基于这三种来源的互联网公司，而谷歌、亚马逊等则是包含三者，全面发展的大数据公司。

大数据时代的管理变革则分为大数据的风险和自由度掌控。大数据带来的风险主要包括个人生活的监视、隐私的泄露、预测惩罚以及数据独裁，对此，通过一定的管理变革，比如个人隐私保护从个人许可到让数据使用者承担责任，预测分析并不决定个人动因及相应惩罚，避免数据独裁而培养大数据算法师以及反数据垄断大亨等，来限制大数据的发展不超过我们可以控制的范围。

引用周涛序中的话对全书内容做一个简要概述，“首先，作者抛出了大数据时代处理数据理念上的三大转变：要全体不要抽样，要效率不要绝对精确，要相关不要因果；接着，从万事万物数据化和数据交叉复用的巨大价值两个方面，讲述驱动大数据战车在材质和智力方面向前滚动的最根本动力；最后，作者冷静描绘了大数据帝国前夜的脆弱与不安，包括产业生态环境、数据安全隐私、信息公正公开等问题”。

**3 感悟思考**

书中对大数据的价值其实也有所分析，主要是站在数据处理角度，分析数据的再利用、重组数据、可扩展数据、数据的折旧值、数据废气、开放数据等带来的数据创新价值，而我的思考方向则是从互联网行业角度出发，分析开放平台、社交网络、电子商务等利用大数据后所获得的改革优势。

**（一）大数据与开放平台**

当我们在人人网上看“最近谁最关心你”，当我们在微博上玩拼图工坊，当我们在app store里下载游戏应用时，我们已经很近距离地接触了当下比较热门的几个开放平台，除此之外还有微信、开心网、淘宝网、百度搜索等等也是很热门的开放平台。

那么，大数据时代到来之际，开放平台面临着怎样的机遇与挑战呢？

开放平台的核心价值是分配流量和消费资源给开发商和应用，同时提高平台本身用户的活跃度和粘度，怎么把平台的流量和用户更好地分配给第三方应用呢，把平台效率最大化是一个巨大的挑战，而大数据是解决这个挑战的重要手段。

开放平台收集了大量的用户数据，而这些数据并不是在完成用户资料系统后就变得一文不值。平台可以根据用户资料系统以及用户浏览了哪些网页、停留了多久、鼠标光标停留的位置、输入了什么信息等“数据废气”实现数据的再利用，了解用户的习惯喜好。数据再利用的价值对那些收集或控制着大型数据集但目前却很少使用的机构来说是个好消息。用户留下的大数据对人人网本身来说，还是停留在好友推荐、智能排序、广告展位等基础优化上，而其能够收集到的大量用户喜好数据对第三方开发商的开放方向和内容则具有很大的参考价值。

随着大数据的出现，数据的总和比部分更有价值，当我们将多个数据集的总和重组在一起时，重组总和本身的价值也比单个总和更大。由于开放平台在信息价值链中的特殊位置，能收集流向每一个API接口的用户数据，寻找各个API接口数据之间的联系，通过数据重组寻找数据间联系，创造数据价值。平台通过数据再利用和数据重组所获得的数据分析结果，可以预测用户的喜好，决定对用户的应用推荐种类和先后关系，对平台流量和用户的流向起到一定的引导作用，有助于实现效率的最大化。

**（二）大数据与社交网络**

人人网、朋友网专注于校园社交，新浪微博、腾讯微博决战140字，米聊、微信激烈争抢移动互联网，还有陌陌、遇见等基于地理位置的社交应用，美丽说、蘑菇街等基于购物分享的社交网站，我们被一大堆社交网站包围着，而一个人能兼顾的社交网站数量毕竟不多，所以他们就成了被选择的对象。随着“社交”理念的逐步渗透、广泛融合，如何提高用户活跃度和粘度，丰富社交网站内容，形成同类网站的竞争优势，是SNS网站能够在竞争激烈的环境下生存下来最重要的砝码。而大数据可以助其一臂之力。

大数据的价值体现在可以挖掘用户的行为习惯和喜好，在凌乱纷繁的数据背后找到更符合用户兴趣和习惯的产品和服务，并对产品和服务进行针对性地调整和优化。比如新浪微博的智能排序功能，根据用户的关注、标签和微博内容等相关信息，帮助用户梳理微博内容，对同类微博进行合并，对可能感兴趣的微博内容进行优先展示，就是利用大数据挖掘用户的关系和喜好，提高用户体验。

2013年4月18日在北京中关村举办的主题为“社交媒体下一浪”的科技交流活动邀请了李开复和youtube创始人陈士骏参与交流。李开复指出，信息爆炸是一个巨大的问题，所以用户需要看跟他个性化最相关的问题，里面非常核心的技术就是如何掌握大数据，如何用大数据经过机器学习达到推进引擎，从海量的数据里提炼出基于用户过去的习惯行为、关注、朋友圈以及发布或者转发的东西，让用户看到最想看到的。可见大数据推动了社交媒体的下一步重要发展。

书中“数据化”一章讲述了随着信息技术的变革，文字、方位、沟通乃至世间万物都可以数据化，也就是数据的丰富度。社交网络的兴起，大量的UGC（User Generated Content，即用户生成内容）内容、音频、文本信息、视频、图片等非结构化数据的出现都依赖于大数据时代的技术发展。而这些信息，恰恰可以成为用户之间建立关系或产生交流的催化剂，从而提高社交网站的用户活跃度和粘度。

社交网络越来越成为网民生活中不可或缺的一部分，而用户也更容易在社交网站上表露自己的身份信息和习惯喜好，有利于社交网站聚集有价值信息的同时，也会带来更大的数据隐忧。首当其冲就是隐私的泄露，即使我们在新浪微博、豆瓣等社交网络上用匿名的方式企图保护自己的隐私，但是在大数据时代，随着数据量和种类的增多，大数据促进了数据内容的交叉检验，又让我们的隐私无处逃遁。同时，互联网的出现，使得对网民的监视更容易、成本更低廉也更有用处。社交网络收集的数据相较于搜索引擎、视频网站、门户网站等其实更贴近用户生活，其大数据战略也就要更注重对用户隐私的保护，不要让大数据发展超出能控制的范围。

**（三）大数据与电子商务**

电子商务本质上是一种销售模式，与线下相比它具有更容易获取消费者数据、商品数据的特点，天猫双十一那天，就有2.13亿独立用户访问，超过1亿笔订单，191亿销售额，这些数据并不只是衡量网站当日的销售情况之后就可以功成身退，而是可以通过用户的购物以及其他行为，分析用户喜好，实现个性化页面展示和推荐，提高广告转化率。同时，类似京东商城等自建物流的B2C电子商务网站，通过相应的大数据分析，有利于合理安排仓库库存，调配物流路线，在效率提高和成本节约上发挥重要作用。

书中举到的很有意思的一个例子是，亚马逊从一开始就已从每个客户身上捕捉了大量的数据，比如他们购买了什么数据？哪些书他们只浏览却没有买？他们浏览了多久？哪些书是他们一起买的？通过数据分析发现，喜欢海明威作品的客户许多也喜欢菲茨杰拉德的书，所以，亚马逊总会将向浏览会购买海明威作品的客户推荐菲茨杰拉德的书，从而增加了销量。如今，据说亚马逊销售额的三分之一都是来自它的个性化推荐系统。可见，大数据创造了一种新的销售模式，主动引导用户增加购买。

我们在对购物用户群体进行分析时，经常会将其分为三类，第一类是购物目标明确，情况知道自己要买什么东西；第二类是购物目标模糊，知道自己要买的商品类别，具体的品牌、外观、价格等特性并不全部都清楚；第三类是没有购物目标，只是以一种随便逛逛的心态，进入购物网站，看到喜欢的也会考虑购买。而大数据战略可以帮助电子商务网站利用用户之前购物所留下的信息，包括曾经购买过什么，收藏过什么，浏览过什么等信息了解用户的喜好，向其推荐可能喜欢的商品，可以增加第一类用户的额外消费，同时简化后两类用户的寻找过程，提高流量的转化率。

同时，对电子商务企业来说，对大数据的充分利用，就可以探索进行个人化、个性化、精确化和智能化地进行广告推送和产品推广，有利于进行更便捷的用户调研，更精准的产品营销，以及更完善的售后监测。

**4 过程体会**

其实大学以来读的书并不多，之前读了克莱·舍基的《认知盈余》，书很好，但至今也未形成书面的阅读报告。大数据是近期来一直在关注的领域，前几周收集了些内容整理成一份报告，小范围内也做过一次演讲，看《大数据时代》一书是在我大数据长期跟进计划中的，这份阅读报告当然也是计划之内的一部分。互联网的趋势，身为计算机学院的学生，自觉是需要有所了解的，所以不论是这次的读书报告，还是之前的内容整理和演讲，都希望自身可以给他人输出价值，共同进步。在本文的最后，想谈谈这次阅读过程中的体会，分享自己的读书过程和读书方法，希望对大家有所参考借鉴之处。

**（一）不只读一本书**

虽然这篇读书报告是我读了《大数据时代》这本书之后写的，但是，读书前后，还查阅整理了很多互联网上的资料，比如谢文的博客，大数据领域相关学者的微博以及许多科技博客中关于大数据的内容，下载观看了许多有关大数据名家对话的视频，比如特邀本书作者维克托·迈尔-舍恩伯格，译者周涛以及谢文、车品觉等大数据领域的知名人士参与的《对话》，李开复、陈士骏对话的节目《社交媒体的下一浪》等，才形成了这样一篇读书报告。

一本书的内容是局限的，即使作者学识渊博，由于篇幅限制也不能把自己所知道的一字不漏地表达出来，所以，看书，要以一本书的内容加以扩展，沿着内容寻找相关的文章、视频，或者沿着作者查阅他之前相关的书或者文章，就是说，要把一本书，当成是钥匙，去开启相关领域的一扇大门。

**（二）先看目录**

拿到一本书，我总是先看目录，本书的目录在之前做大数据报告时也看过，当时未细看，现在重新好好看一遍才理解其中的逻辑关系。

整本书分为三个部分，分别是大数据时代的思维变革、商业变革和管理变革，对大数据的观念发生改变，应用实践才能实现大数据的价值，而在大数据商业化的过程中，又要注意对其管理统筹，实现大数据价值的最大化。

看目录有利于对整本书整个的逻辑思路有一初步的了解，同时在阅读的过程中，对目录的熟悉也会对读者思考起到一个承前启后的作用。

**（三）序看两次**

序其实是书本很精华的一部分，一般由同一领域比较权威的人执笔，对整本书的内容和价值进行了一定的介绍，专业度也比较高。由于序的篇幅较短，信息量大，一般是原理性内容偏多，缺少实例来支撑其原理，内容比较难懂而且可能还有点枯燥，所以我之前对于序的做法一直都是跳过，这次看书，我在看书前和看书后将序看了两遍，看书前是为了对书的内容有一定的了解，看书后是对书中内容的一次回顾整理。

参考文献

[1]维克托.迈尔—舍恩伯格,肯尼思.库克耶. 大数据时代[J]. 浙江人民出版社.

[2]任磊,杜一,马帅,张小龙,戴国忠. 大数据可视分析综述[J]. 软件学报,2014,09:1909-1936.

[3]程学旗,靳小龙,王元卓,郭嘉丰,张铁赢,李国杰. 大数据系统和分析技术综述[J]. 软件学报,2014,09:1889-1908.

[4]王元卓,靳小龙,程学旗. 网络大数据:现状与展望[J]. 计算机学报,2013,06:1125-1138.