第 5 章　算法，让数据有了价值

今天的人们早已从信息匮乏的时代进入了数据大爆炸的时代，每个人都是信息消费者，也是信息制造者。然而，我们却越来越难以从大量信息中找到自己感兴趣的内容，也越来越难以让自己生产的信息脱颖而出，被别人看到。在数据的汪洋大海中，并不是每一朵浪花都具有价值，可我们却常常被那些无用的信息裹挟，深陷其中，迷失方向。数据本身并不产生价值，按照大数据专家的观点，首先需要挖掘数据中的信息，然后从信息中建立结构并发现知识，再把知识转换成能力，基于能力建立系统，系统中策略的执行才最终产生价值。这个链条可真长！我们可以简单理解为只有有序的数据，能够为用户提供有用信息的数据才能产生价值。那么到底如何才能从数据中找到价值呢？在赛博的世界里，再也没有谁比算法更有能力给数据赋予价值了。算法是如何做到这一点的呢？

大数据的发展带来了人类新文明

计算机总是越来越智能的。科学家告诉我们不久它们就能跟我们对话了（这里的“它们”，我指的是“计算机”。我怀疑科学家永远都不能跟我们对话）。

——达沃·巴里，作家

每个时代都有自己的“大数据”

如果我们回顾历史，会发现每一种文明的更替都伴随着信息载体的变迁，而信息载体的变迁，又带来了信息量的增长。无论在哪个时代，都存在着对于那个时代的信息载体和信息处理工具来说是“规模很大”或“很复杂”的数据集合。从人类诞生之日起，数据就不断产生，例如部落的人口、捕杀的猎物、农田的收成，甚至口口相传的传说都可以视为数据。数据逐渐产生、积累，导致数据规模越来越大，数据之间的关系也越来越复杂。当数据规模和复杂程度逐渐达到甚至超出人类当时的数据处理能力后，即可被视为“大数据”——尽管当时的人们并没有“大数据”的概念。

在文字诞生之前，原始先民生活中发生的事件只能靠个人头脑记忆。后来发生的事情越来越多（即“数据”规模越来越大），只靠记忆难免发生混乱。为了避免忘记以往的事情，先民们需要花上一整天的时间来复习过去发生的事情，然而悲催的是，每天都有新的事件发生。这无疑使得原本困难的生活雪上加霜，记忆和回忆占用了大量时间，都没有足够的时间进行渔猎填饱肚

子了！

幸运的是，传说中神农氏发明了结绳记事，拯救万民于水火，使得先民不必每天花费大量精力记忆众多的琐事。《周易·系辞下传》中说："上古结绳而治，后世圣人易之以书契。"其意思就是，上古的人们通过在绳子上打结来记事，后世的圣人发明了文字，开始用书写和文字来代替结绳记事。东汉经学家郑玄在所著的《周易注》中解释说："古者无文字，结绳为约，事大，大结其绳；事小，小结其绳。"《周易集解》进一步解释说："古本无文字，其有约誓之事，事大，大其绳，事小，小其绳，结之多少，随物众寡，各执以相考，亦足以相治也。"图 5.1 给出了用于记事的绳结示意图。

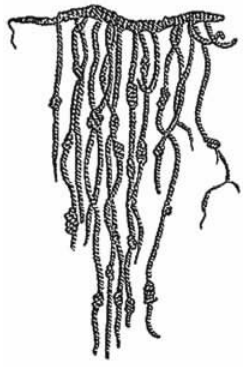

图 5.1　用于记事的绳结

结绳记事虽然方法简单，实际上也包含一套相对完备的"算法"，即大事对应大绳结，小事对应小绳结，数量多就多打绳结，数量少就少打绳结，这套简单的机制在原始社会中"足以相治"，可以满足一般的生活需求。结绳记事这种新的"数据处理"方式，配合语言交流，使人们可以更加方便地记住更多的事情，应对了那个时代的"大数据"挑战。与此类似的还有农耕文明中的甲骨文和封建文明的造纸术。

近年来，随着互联网和移动通信技术的发展，全球数据总量呈爆炸性增长，增长速度也逐年加快。1984 年诞生于美国旧金山的思科(Cisco)公司，是全球 IT 企业中的巨无霸。作为全球顶级的网络解决方案提供商，仅 2016 财年第四季度的收入就达 126 亿美元。思科公司在《思科可视化网络指数：全球移动数据流量预测白皮书(2015—2020)》中，对 2015 年全球的移动数据流量做了详细的统计分析，并对未来五年的移动数据流量的增长做出了预测，如图 5.2 所示。与 2014 年相比，2015 年全球移动数据流量增长了 74%。另

外一个令人震惊的数值是，2015 年全球移动数据流量是 15 年前的 4 亿倍，是 10 年前的 4000 倍。截至 2015 年底，全球每个月产生的移动数据量平均为 3.7EB，而在 2014 年底时，这个数值是 2.1EB。据思科公司预计，到了 2020 年，全球月均移动数据流量将增长 8 倍，达到 30.6EB。这个数据量堪称惊人。要知道，1EB 数据包含 1 152 921 504 606 846 976 个字节，存储 1EB 的数据需要 1 048 576 块 1TB 的硬盘。

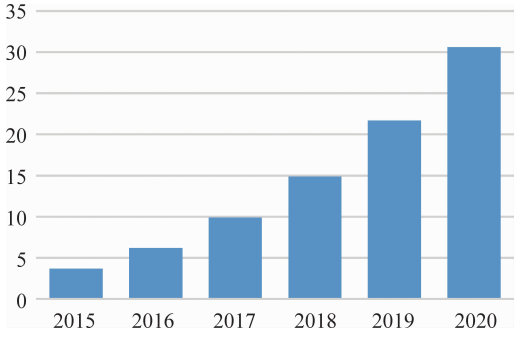

图 5.2　思科公司预测 2015—2020 年间全球移动数据流量增长情况（单位：EB）

全球著名的科技咨询顾问提供商国际数据公司（International Data Corporation，IDC），服务领域主要集中在信息技术和电信等行业。IDC 在报告 *THE DIGITAL UNIVERSE IN* 2020：*Big Data*，*Bigger Digital Shadows*，*and Biggest Growth in the Far East* 中指出，2020 年的全球数据总量将是 2005 年的 300 倍，达到 40 000EB；并且从 2013 年开始，全球数据总量将 2 年左右就翻一番。根据专家的预测，到 2020 年，每位互联网用户平均每天要产生 1.5GB 的数据流量，一辆无人驾驶汽车每天产生的数据量将达到 4000GB，一架飞机每天产生的数据量将达到 40 000GB，一座智慧工厂每天产生的数据量更是超过 1 000 000GB。

大数据，实际上就是人类文明的产物。不同的时代，因为数据处理能力不同，对何种数据集合是“大数据”也有不同的认识。过去的大数据，可能就

是现在的“小数据”;现在的大数据,可能就是将来的“小数据”。大数据永远是一个相对的概念。

运转不停的世界每天都有大量数据产生,未来还将有更加大量的数据产生。并且,在信息技术的催动下,数据洪流以前所未有的速度席卷到我们生活中的每个角落,我们每个人都身处其中,没有人能够例外。

赛博时代的大数据

虽然每个时代都有自己的大数据,但是大数据真正被我们熟知却并没有多久。一般认为,20 世纪 90 年代,由于计算机科学家约翰·马西(John Mashey)的大力推广,大数据一词才慢慢流行起来。我们能从图 5.3 的百度指数清晰地看出 2011—2016 年以“大数据”为关键字的搜索趋势(上面曲线)和媒体关注(下面曲线)变化情况。图 5.4 是谷歌趋势给出的 2004—2016 年以“Big Data”为关键字的搜索热度。从这两个图都可以看出,大数据一词大概于 2012 年才慢慢进入公众视野,并成为一个越来越热门的话题。

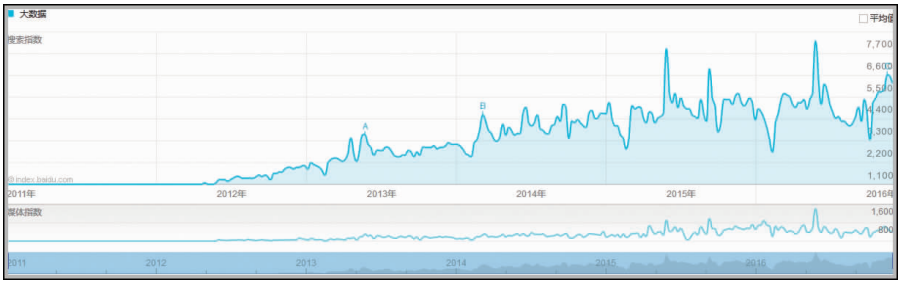

图 5.3 “大数据”一词的搜索指数和媒体指数

根据维基百科的定义,大数据(Big Data),又称巨量数据、大资料,指的是所涉及的数据量规模巨大到无法通过人工或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。百度百科则将大

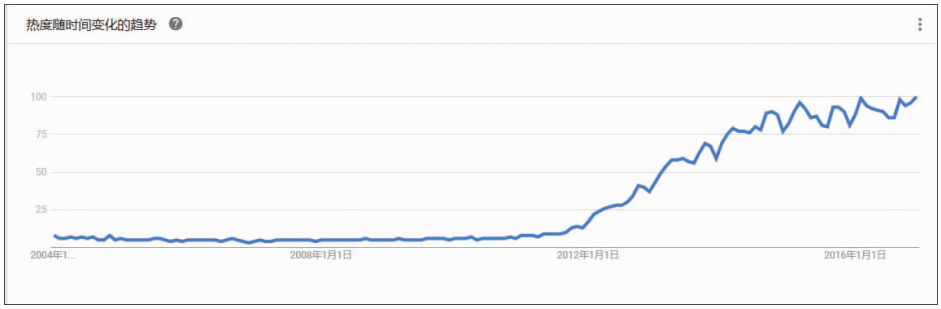

图 5.4　“Big Data”一词的搜索热度

数据定义为：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，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。

这听起来也许有些抽象。如果不使用这些严谨的科学术语，大数据到底是什么呢？大数据和普通的数据有什么区别呢？其实，通俗地说，如果数据规模不断增长或者复杂程度不断提升，以至让我们觉得这些数据处理起来很棘手时，普通的数据就一跃成为“大数据”。

大数据一词流行之前，人们通常使用海量数据（Mass Data 或 Massive Data）一词来指代大规模的数据集合。但是，我们不应简单地把“大数据”当作“海量数据”的同义词。事实上，赛博时代的大数据除了数据规模大之外，通常还具有流转速度快、数据类型多、价值密度低等特点，这就与以往“海量数据”有了本质区别。

大数据蕴含的大能量

人类活动产生了大量的数据，这些大数据中存储了大量的信息，大数据对人类的价值就在于这些信息的价值。我们先从一件小事中一窥大数据蕴

含的“能量”。

1986年，美国自动化领域的先驱、迪堡集团有限公司（The Diebold Group，Inc.）的创始人约翰·迪堡（John Diebold）提到了这样一件趣事。1979年，一家银行在某地区安装了ATM机网络。但是银行发现，某一台取款机每天凌晨0点到2点之间都会发生大量的提款操作。这是怎么回事呢？银行怀疑这里发生了违规行为，于是专门雇用了侦探对此进行调查。调查的结果充满戏剧性，侦探发现这个提款机刚好位于红灯区附近，午夜光顾红灯区的人们当然不想刷卡消费，因为这会在信用卡交易中产生一笔“不光彩”的记录。于是，这些顾客总是先在这部ATM机上取出现金，然后去红灯区消费。

如果通过一台提款机的取款记录能算出某个特定区域的消费特征，那么如果把全国所有提款机的取款记录都做采集分析，再辅以特定的计算机算法，能算出的就不仅限于红灯区消费了，甚至可以更详细地了解全国人口的储蓄能力、消费水平、消费习惯、资金往来、资产负债等情况。

大数据还可以让搜索引擎在公共卫生领域发挥作用。谷歌公司曾于2008年推出了流感趋势系统（Google Flu Trends）和登革热趋势系统（Google Dengue Trends），这套系统采集了美国人平时经常使用的5000万个搜索关键字和2003—2008年间美国疾控中心公布的疾病传播数据，发现了特定搜索关键字和疾病传播数据之间具有相关性。如果某地区包含特征字段的搜索关键字数量有异常上升，就可以认为该地区已经开始发生疫情传播。

为了充分挖掘大数据中存储的有效信息，谷歌公司共测试了4.5亿个数学模型，并从中确定最有效的预测模型。通常，美国疾控中心（Centers for Disease Control and Prevention，CDC）公布的疫情传播报告通常比实际传播情况滞后两周。因此谷歌公司声称，这套系统可以比CDC更早获知流感疫

情，并且该系统预报的疫情和官方公布的数据相关性高达 97%。虽然也有专家对此数据表示质疑，但不可否认的是，大数据技术正在以我们意料不到的速度和方式渗透到我们的生活里，为研究人类行为和人与人之间大规模的互动提供了崭新的思路。

2015 年 1 月 25 日，朋友圈发生了一桩“灵异”事件——一些人发现“宝马中国”在他们的朋友圈发了一条状态；一些人的朋友圈出现了“vivo 智能手机”的文字和图片，而另有一些人的朋友圈则出现了“可口可乐”。很快真相大白，原来在这一天，已经积攒了 5.49 亿用户的微信，正式启动了基于朋友圈大数据分析的新型广告推送（图 5.5）。

图 5.5　微信朋友圈的第一条广告

与传统广告全面铺开的传播方式相比，微信试图通过用户大数据分析，把广告推送给可能会感兴趣的用户。通过用户的注册信息、绑定的手机号和 GPS 定位，可以判断出用户所在地，通过朋友圈里分享的帖子，则可以分析出用户的职业特点、兴趣爱好等，钱包支付记录则体现出用户的消费水平，卡包里各种会员卡体现出用户的品牌倾向，除此之外，在朋友圈的点赞、留言也都可以体现出用户的个性和偏好等。这些庞大的用户行为数据如同一座宝藏，微信可以通过不同的算法了解到用户的特点和需求，从而推算出用户感兴趣的东西。

算法，点石成金的力量

你把数据拷问到一定程度时，它自然就会坦白一切。

——罗纳德·科斯，诺贝尔经济学奖获得者

大数据并不等于大价值，就像开采矿藏一样，没经过开采的数据仅仅是一堆数据。而借助算法，才能挖掘出大数据中熠熠生辉的珍宝。算法，赋予了大数据生命，让数据可以为我们所用。

机器学习算法——计算机会学习

随着数据规模的增大和复杂程度的提升，人类自身的力量变得越来越渺小，如此规模的数据显然是人力难以处理的，所以只能求助于计算机和数据处理算法。实际上，在大数据面前，简单的算法也往往因为效率太低而难以奏效，于是人们寄希望于计算机可以自己学习大数据中存储的信息，这就是机器学习。

从“机器学习”这个名字不难看出，这门科学研究的就是如何使机器具备学习能力，进而掌握知识具备智能。可以说，机器学习的终极目标就是人工智能。机器学习的产生和发展与人工智能技术的发展息息相关，其应用范围也主要集中在人工智能领域。

这两年人工智能步入了蓬勃发展的阶段。但是，人工智能从诞生到现在，也经历过三起三落，既获得过人们热切的期盼，也经历过漫长的寒冬。人工智能的历史，是一部群星璀璨、高潮迭起的历史。

作为一门科学，人工智能诞生于 1956 年的达特茅斯（Dartmouth）会议。这次会议上，相关领域的科学家就这门学科的名称达成一致，即 Artificial Intelligence（AI），并且明确了人工智能研究的任务。

人工智能自诞生以来，大致经历了三个阶段，分别是推理期、知识期和学习期。

从 20 世纪 50 年代中期到 70 年代初期，可以称为人工智能的“推理期”。由于达特茅斯会议的影响，这一阶段人工智能得到了大量资金投入，受到了业界、政府和学术界的广泛关注，这是人工智能发展的黄金时期。然而，就如同大多数新技术成长过程往往充满曲折一样，人工智能也是如此，人们慢慢发现他们当初的乐观预期迟迟无法实现。于是，20 世纪 60 年代后期，人工智能研究渐渐步入低谷。

20 世纪 70 年代中后期到 80 年代中期，是人工智能发展的第二个阶段——“知识期”。当时，以爱德华·费根鲍姆（Edward Feigenbaum）为代表的科学家提出，让机器拥有智能的前提是使机器拥有知识。于是，“知识处理”和“知识工程”成为人工智能的主流研究方向，“专家系统”（Expert System）开始被大家接受。专家系统是具备专业知识的计算机智能程序，可以扮演领域专家的角色。专家系统带动了人工智能的再次繁荣。1981 年，日本启动了以面向知识处理为主要目标的“第五代计算机”计划。然而，随着研究深入，人们发现知识的表示是一件很困难的事。于是，知识工程进入瓶颈期，“第五代计算机”预先设定的目标也没能按期实现。与此同时，由于世界范围内经济泡沫破裂，人工智能再次进入低潮，进入了人工智能的“寒冬期”。

20 世纪 90 年代后期，人工智能进入“学习期”。知识的表示是一个难题，那么能不能让机器自己学习知识呢？在这个阶段，机器学习融合了符号主义和联结主义，成为人工智能研究主流，产生了很多让人眼前一亮的成果。目前，机器学习已经成功应用到计算机视觉、自然语言处理、搜索引擎、语音