

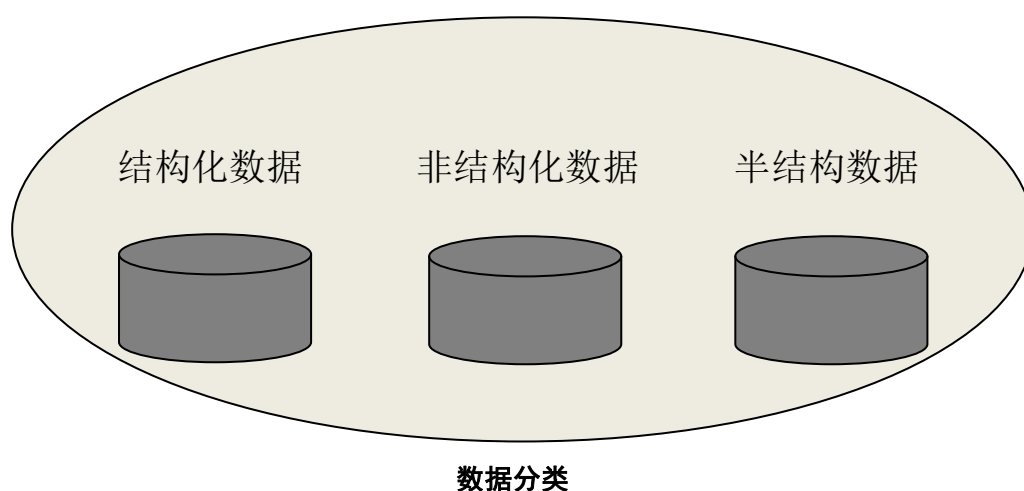
大数据简介

林才贵

重庆邮电大学

一、什么是大数据？

谈起大数据，很多人都不陌生，不管是专业，非专业人士都耳熟能详。一开始这个“大数据”并不大，互联网还没有兴起的时候，数据也就是我们读过的书籍、报刊等，一个星期加起来的书籍才有多少字节？如果你不是生活在大城市，一个普通学校图书馆的书加起来也没有多少信息，后来随着信息化社会的到来，互联网的到来，才让数据爆炸式增长。首选我们抛开“大数据”的“大”，来看看数据都有什么。数据分为三类，一种叫结构化数据，一种叫非结构化数据，还有一种叫半结构化数据。



结构化数据：即固定格式和有限长度的数据。例如填的表格数据，姓名：xxx，名族：汉，性别：男。这都叫结构化数据。

非结构化数据：现在的非结构数据越来越多，比如网页，有的网页很长有的很短；还有就是语音、视频、图片这些都是非结构化数据。

半结构化数据：是一些 XML 或者 HTML 的格式，没有从事技术的人可能对这部分数据不太了解，不过没有关系，我们只需要知道这部分数据是介于结构化和非结构化数据之间的数据，即部分格式化和有限长度的数据，部分没有格式化和固定长度的数据。

数据为何大起来？换言之，前面提到的三类数据为何突然之间就多起来了？其实也不是突然之间，而是得益于计算机硬件的发展和互联网的发展。我们都知道我们用的手机内存卡从原来的多少 M 到现在的多少 G，最直观的就是，每次各大手机厂商的新品发布会，都会把多少 G 运行内存加多少 G 的内存拿来“秀一把”，因为内存的大小决定了你处理多少数据的

快慢和存储多少数据。存储数据的介质在不停的发展，而传播数据的介质同样发展迅速——互联网。互联网将全世界的数据链接在一起，相当于一个超级大的磁盘，存储着超级多的数据。只要你愿意，你就只需要敲一下键盘或者点击一下手机屏幕，存储在你磁盘里面的数据就可以分享给世界，同样你也可以拿到别人分享的数据。



大数据 (图片来源于互联网)

二、大数据时代的“四V”

自互联网的诞生以来，数据就一直以惊人的速度增长。门户网站、搜索引擎、购物网站、社交软件使得数据不断的膨胀。智能移动端的流行让数据如虎添翼，智能手机每天都在收集用户的信息，比如位置、运动轨迹、生活信息等数据。以及 5G 时代推动着物联网进一步革新，而物联网又要依赖于各种传感器，传感器收集的数据会以爆炸式的方式增长。

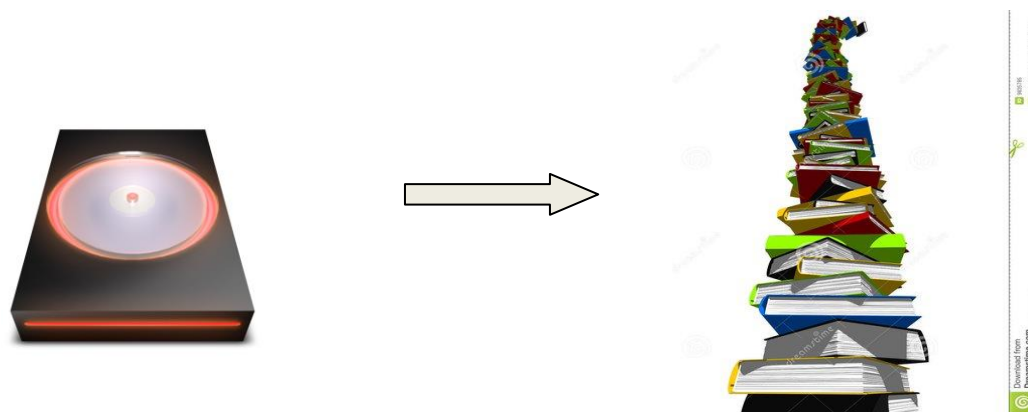
“大数据”已经成为最近几年最火词汇之一，同等“人工智能”、“云计算”。

IBM 提出“三V”的概念，即大量化 (Volume)、多样化 (Variety)、快速化 (Velocity)，是大数据时代的显著特征。而大数据时代的“四V”是在“三V”的基础上增加了第四个字母为V的词—Value (价值)。这四个V无非是从各个方面来描述大数据的特征。

在大数据的四“V”中，大量化 (Volume) 是显而易见的。如果没有大量数据，我们就无法称其为“大数据”。如今，各企业的数据正在从 GB、TB 级别向着 PB、EB 级别迈进。1PB = 1,024TB = 1,048,576GB = 1,125,899,906,842,624Bytes (字节)，而 1EB=1,024PB=1,048,576TB。

下面举个简单的例子，以 GB 为单位换算成图书：

- (1) $10\text{GB} = 40 \times 1000 \times 1000 \times 1000 = 10,000,000,000\text{B} = \text{一百亿字节}$ 。
- (2) 1 个汉字需 $1 \times 2 = 2\text{B}$ ，10 万字需 $100,000 \times 2 = 200,000\text{B} = \text{二十万字节}$ 。
- (3) $\text{一百亿字节} / \text{二十万字节} = 10,000,000,000 / 200,000 = 50,000$ 本
即五万本。
- (4) $1 \text{ 厘米} \times \text{五万本} = \text{五万厘米} = 500 \text{ 米}$ ，世界最高的建筑哈利法塔也不过 800 米。



多样化（Variety）是指结构化、非结构化、半结构化数据一起在飞速发展。全世界四五十亿手机用户成为数据提供者，同时手机中的各种传感器使得非结构化数据正在以一定增长速率赶超结构、半结构数据。



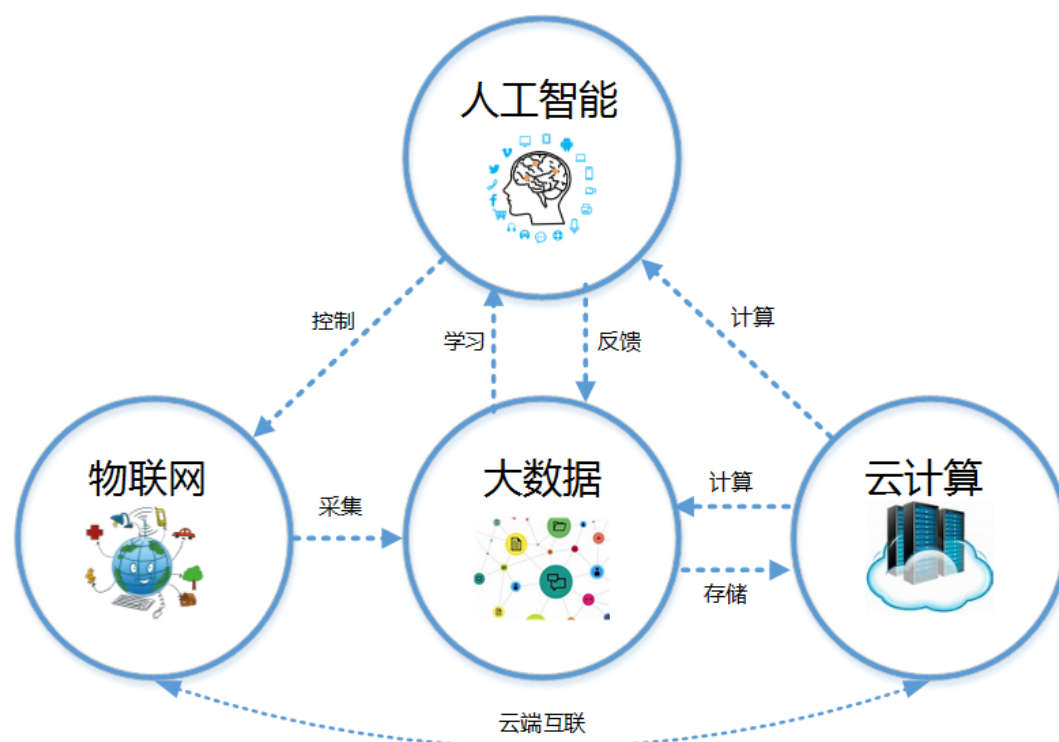
数据多样性 (图片来源于互联网)

快速化（velocity）主要是指商业和各个相关领域处理的交易以及数据在以越来越高的速度和频率产生。每一分钟就有大量的数据在商业环境和互联网环境中产生。

价值（value），则是指数据运营和应用的重要性。所以就催生出数据分析，数据挖掘这样的课程和工作。数据还只是数据。只有通过处理和分析过的数据才能转换为信息，归纳成知识。“知识改变命运”我们都知道。曾经业内出现过“啤酒和尿不湿”的成功案例，沃尔玛超市通过数据分析发现，将啤酒与尿不湿摆放在一起，同时将牛肉干等一些简便的下酒食品也摆放在一起，这样全年下来，营业额增加了几百万美元。原因是因为啤酒和尿不湿多为男人在周末采购，而且购买这两种产品的顾客一般都是年龄在25至35周岁的青年男子，由于孩子尚在哺乳期，多数男人都接到了夫人的“圣旨”，下班后带尿不湿回家，而周末正是美国体育比赛的高峰期，一杯啤酒伴着比赛是多么惬意的事。简单的数据分析，就变成了财富，所以数据的价值（Value）才是最重要的。

三、大数据和云计算、人工智能的关系

数据大了，就会想办法怎么存储下来，而存储下来的数据又要将其转化为知识，就需要各种方法来实现。一台机器很难存储处理这些数据，于是有人就想到了把大数据拆分成若干小数据，同时用很多机器来存储处理。云计算就是把很多设备连接起来，抽象各种资源，比如计算资源、存储资源、网络资源，使他们统一起来解决大数据的存储和计算问题，当然还有其他功能，这里我们就只提及云计算与大数据的关联。人工智能就是使用各种算法（方法）找到大数据背后的规律，然后将规律应用。我们熟知的自动驾驶技术、打败柯洁的 AlphaGo、人脸识别技术等通过大量的数据用人工智能方法比如统计学、机器学习，还有目前比较火热的深度学习来找规律，实现智能。大数据、云计算、人工智能的关系非常的密切，相辅相成，谁也离不开谁。



大数据、人工智能、云计算、物联网、关系图

(图片来源于互联网)

四、大数据未来发展

由于现今数据量级猛增以及人们对数据中潜在信息的重视，数据挖掘的研究会受到越来越多的关注，在最近几年更是有了长足的进步，如以下几个方面：

- (1) 对于大规模数据的存储、管理和使用，包括在分布式环境建立数据仓库的方式方法。
- (2) 知识发现语言的形式化描述和算法，即研究专门用于知识发现的数据挖掘语言。
- (3) 数据挖掘过程中的可视化方法，使知识发现的过程中能够更容易被用户理解，也便于在知识发现的过程中进行人机交互。
- (4) 生物信息或基因的数据挖掘。
- (5) Web 数据挖掘的各个方面。

(6) 因为 5G 技术的全面应用, 会促进物联网的进一步发展, 在万物互联的未来, 各种设备产生的非结构化数据将会成为重点研究的对象。

这些数据—原始石油, 需要提炼出我们需要的汽油、柴油等。如何发现这些数据背后的知识, 是我们每个人都想知道的, 它就像一座金矿, 而我们就是拿着各种工具的矿工, 迫不及待的想去发现金子。

作者简介: 林才贵, 重庆邮电大学硕士研究生, 研究方向: 分布式智能系统。

文章参考文献

- [1] 谭 磊. 大数据挖掘[M]. 北京: 电子工业出版社, 2013.3: 1-11.
- [2] 刘超. 不是技术也能看懂云计算, 大数据, 人工智能
[EB/OL]. www.cnblogs.com/popsuper1982/p/8505203.html.
- [3] Bruce Ratner, Ph.D. Statistical Modeling and Analysis for Data(4th printing) [M]. CRC Press, ISBN:1574443445.
- [4] Rahimi H, Zibaeenejad A, Safavi A A. A Novel IoT Architecture based on 5G-IoT and Next Generation Technologies[J]. 2018.
- [5] Obschonka, Martin & Audretsch, David. (2019). Artificial Intelligence and Big Data in Entrepreneurship: A New Era Has Begun. Small Business Economics. 10.1007/s11187-019-00202-4.
- [6] Zhao Y, Calheiros R N, Gange G, et al. SLA-Based Resource Scheduling for Big Data Analytics as a Service in Cloud Computing Environments[C]// 2015 44th International Conference on Parallel Processing (ICPP). IEEE, 2015.
- [7] Naimi A I, Westreich D J. Big Data: A Revolution That Will Transform How We Live, Work, and Think[M]. 2013.
- [8] Wu X, Zhu X, Wu G Q, et al. Data mining with big data[J]. IEEE Transactions on Knowledge & Data Engineering, 2013, 26(1):97-107.



临菲信息技术港



临菲信息技术港 公众号



临菲学堂