第 1 章　基础知识

1.1　模拟和数字化

现实世界中我们看得见、摸得着的物品经常使用模拟信息表示其属性，如物品的长度、高度和宽度。模拟信息最重要的一个特点是连续性，即在某个区间产生的连续值，如桌子的长度是 2.15 米。这个模拟信息值仅是一个相对准确的概念，或者说是一个近似值，因为桌子的长度往往不是恰好 2.15 米，而是近似 2.15 米，这主要取决于测量工具的精度。测量值小数点后的位数随着测量工具的精度增加。模拟信息的另一个重要特点是无限性。科技的进步让测量精度可以增加到非常多，甚至无限多的小数位数。在模拟世界中可以借助某种设备用测量的方法取得模拟信息的数值，数值是一个无限小数，介于两个相邻的数值之间，这两个相邻的数值随着精度的增加可以无限分割。

在计算机和网络世界中，任何数据都使用有限个“0”和“1”组合的代码来表示，如计算机中的数字、文字、图片、声音、视频和动画等数据。美国信息交换标准码（American Standard Code for Information Interchange，ASCII）是计算机最早使用的编码。如字母“A”的 ASCII 编码为“1000001”。计算机系统不存在无限的概念，因为任何数据均存储在有限的内存或外存中，所以存储数据时必须使用有限的位数表示。在计算机系统中，数据最大的特点是离散性，即孤立的点集。如整数集的任何两个元素之间都有一定的距离，任何两个连续的整数之间无任何其他整数值，即任何两个连续的整数之间无法继续分割。

思考：（1）计算机中的小数是离散的还是连续的？[1]

（2）计算机中的颜色是离散的还是连续的？[2]

1.2　数模转换

随着计算机和网络的普及，现实世界的模拟数据需要保存到计算机中，然后经过计算处理后在网络上共享、传播。这个过程需要将模拟信息转换为数字数据，也称为数模转换过程，有些书籍、论文中也称为数字化过程。

---

[1] 计算机中的小数是离散的、非连续的。现实世界中，任何两个连续的小数中间都可以插入其他小数，如 9.25 和 9.26 看似连续的小数，但实际上二者之间可以插入无限个小数，如 9.251、9.2527 和 9.25999 等，插入的小数只要在范围(9.25, 9.26)即可。但在计算机中，由于存储空间的限制，在两位小数集合中，9.25 和 9.26 就是连续的小数，二者中间无法插入其他两位小数。

[2] 计算机中的颜色是离散的、非连续的。现实世界中，颜色是无限的、连续的模拟信息。计算机中的颜色模型有很多种，但每种模型包含的颜色个数都是有限的，任何两个连续的颜色中间不能存在其他颜色。保存在计算机中的任何数据都是离散的。

模拟信息转换为数字数据，需要采样（sampling）和量化（quantization）两个步骤。采样也称为取样或抽样，是将无限的、连续的模拟信息转换为有限的、离散的数据。例如，将时间轴上连续的信号每隔一定的时间间隔抽取出一个信号的样本，使其成为时间上离散的序列。量化是将信号的连续取值近似为有限个离散值的过程。

采样过程中涉及采样率的概念，即抽取信号的时间间隔。量化过程中涉及位深的概念，即量化的等级。为了方便计算机处理，量化一般为 2 的整数次幂。如将纸质黑白图片输入计算机时要进行数模转换，采样率即图片分辨率，量化即为灰度级。将声音输入计算机时，采样率即多长时间的间隔获取一个声音属性，量化即为声音的幅度属性。采样率和位深决定模拟信息转换为数字数据的质量，采样率越高、位深越大质量越好，但存储文件会增大，计算机处理和网络传播都会受到影响。所以，考虑到人眼和人耳的辨识能力、数字文件的大小、计算机的处理能力和网络传播速度等原因，采样率和位深有一个理想的数值，如为获取 CD 音质的音频采样率一般是 44100 Hz，即每秒采样 44100 个；位深是 16 比特，即将声音的振幅分为 65536（2^{16}）个等级。

模拟数据在读取时，由于测量设备的原因，只能是一个近似值。经过数模转换后，很多信息不能被精确地表示，也是一个近似值。

1.3 进制

日常生活中人们接触的大多是十进制数据[3]，“如十两等于一斤”，而在计算机系统中采用二进制表示和处理数据，十六进制存储数据。

十进制包含十个基数，分别是 0，1，2，3，4，5，6，7，8，9。基数的排列组合表示一个数值，基数相同但位置不同表示的数值也是不同的。例如，某单位某天的营业额是 1011 元，4 位数字根据位置代表的数值分解如下：

$$\begin{aligned}1011\text{（十进制）} &= 1\times10^3 + 0\times10^2 + 1\times10^1 + 1\times10^0\\ &= 1000 + 0 + 100 + 1\\ &= 1011\end{aligned}$$

进制中每个固定的位置对应的单位值称为“位权”。十进制的特点是“逢十进一”，位权是 10^n。

二进制包含两个基数，分别是 0 和 1。例如，计算机中的二进制数据 1011 代表的数值分解表示如下：

$$\begin{aligned}1011\text{（二进制）} &= 1\times2^3 + 0\times2^2 + 1\times2^1 + 1\times2^0\\ &=8 + 0 + 2 + 1\\ &=11\text{（十进制）}\end{aligned}$$

二进制的特点是“逢二进一”，位权是 2^n。

十六进制包含 16 个基数，分别是 0，1，2，3，4，5，6，7，8，9，A，B，C，D，E 和 F。其中，A～F 分别表示十进制的 10～15。例如，计算机中存储的十六进制数据 1011 代表的数值分解表示如下：

$$1011\text{（十六进制）} = 1\times16^3 + 0\times16^2 + 1\times16^1 + 1\times16^0$$

[3] 日常生活中非十进制的数据有 60 秒是 1 分钟，7 天是 1 周，12 个是 1 打（dozen），3 尺是 1 米等。

= 4096 + 0 + 16 + 1
= 4113（十进制）

十六进制的特点是“逢十六进一”，位权是 16^n。

数字“1011”在不同的进制中表示的值是不同的。计算机使用二进制是为了技术简单、运算规则简化、安全和可靠等原因。但二进制需要更多的位数存储数据，所以计算机使用十六进制存储数据，使用二进制处理和计算数据。

1.4 存储单位

计算机系统中表示或存储数据的最小单位是“位”（bit，binary digit），也称为比特。如二进制值中的每个“0”或“1”需要 1 bit 存储。在 ASCII 中，字母“a”的编码为 97，使用二进制表示为“1100001”，即字母“a”需要 7 位来表示。

比特单位比较小，所以经常使用字节（Byte，简写为 B）作为数据存储或表示的单位。1 字节表示 8 比特。随着人们对数据的重视程度越来越高，政府、国际组织、公司甚至个人都开始数字化数据，所以需要更多的单位表示日益增加的数据。计算机中经常使用前缀来表示更多的存储单位，见表 1.1。

表 1.1 存储单位前缀

前缀	含 义	举 例
K	2^{10}	1KB = 2^{10} B = 1024 B
M	2^{20}	1MB = 2^{20} B = 1024 KB
G	2^{30}	1GB = 2^{30} B = 1024 MB
T	2^{40}	1TB = 2^{40} B = 1024 GB
P	2^{50}	1PB = 2^{50} B = 1024 TB

1.5 因特网

Internet 即因特网，也称为国际互联网，是世界各地的网络利用 TCP/IP 技术，通过路由器连接而成的覆盖全世界的全球性互连网络，用户只要接入因特网中的任何一台计算机，就意味着已经登录 Internet。Internet 提供了许多重要的服务，常见服务如 WWW、FTP、Telnet、BBS、Mail 等。

1969 年，美国国防部高级研究计划局（Advance Research Projects Agency）出于军事需要组建了阿帕网（ARPANET），这就是现在因特网的前身。1985 年，美国国家科学基金会（National Science Foundation）开始建立以科研和教育为目的的全国性教育科研网（NSFnet），代替了 ARPANET 的骨干地位。1989 年，MILNET（由 ARPANET 分离出来）与 NSFnet 连接后，就开始使用 Internet 这个名称。20 世纪 90 年代初，商业机构进入 Internet。1995 年，NSFnet 停止运作，Internet 被彻底商业化。

1989 年，我国开始建设因特网。1994 年 4 月 20 日，中关村教育与科研示范网络（中国科技网的前身）率先与美国 NSFnet 直接互连，实现了中国与 Internet 全功能网络的连接，标志着我国国际互联网的诞生[4]。

随着人类从工业社会向信息社会的过渡，因特网的发展异常迅猛，政府、公司、团体或个人都可以方便地使用因特网获取、发布和传输信息。因特网已经成为人们日常生活的一部

[4] 数据来源于中国互联网信息中心。

分。截至 2019 年 6 月，中国网民规模达 8.54 亿，中国手机网民规模达 8.47 亿 [5]。

互联网（internet，注意首字母小写）不同于因特网（Internet，注意首字母大写）。互联网是指能彼此通信的设备组成的网络，因特网是互联网的一种。互联网为实现彼此通信可能采用多种协议，而因特网是使用 TCP/IP 实现通信功能的广域网。由于因特网的使用更广泛，在很多领域中将互联网等同于因特网，二者互用的时候很多，均表示因特网。

广域网、城域网和局域网是对网络规模的一种分类方法。局域网（LAN）覆盖的地理范围一般在 10 千米以下，如一个学校或一个公司。广域网（WAN）也称为远程网，覆盖的地理范围为几十到几万千米，甚至横跨一个或多个国家。城域网（MAN）的范围介于广域网和局域网之间，覆盖的地理范围大约是几十千米。

万维网（World Wide Web，简称 WWW 或 W3）是 Internet 提供的一种重要服务，万维网包含无数个网络站点和网页，使用超链接连接多媒体。万维网起源于 1989 年，是为了研究的需要，由 CERN（欧洲粒子物理实验室）的研究人员开发的一种远程访问系统。目前，大多数企事业单位、公司和组织都在因特网上建立了自己的万维网站。

1.6 地址和协议

地址和协议是因特网的两个重要概念，地址用于辨识因特网中的每台计算机，协议用于保障两台计算机无障碍地进行信息沟通。

1．地址

作为因特网的一项重要服务，WWW 可以提供信息的共享和远程访问。为了快速地访问某个 WWW 服务，提供 WWW 服务的计算机需要有一个 IP（Internet Protocol）地址。事实上，连接到因特网上的每台计算机都需要一个 IP 地址。

IP 地址分为静态 IP 和动态 IP 地址两种。静态 IP 地址的特征是每台计算机分配一个固定的 IP 地址。由于 IP 地址的个数是固定的，不能保证每台连接因特网的计算机都有静态 IP 地址，因此动态 IP 地址应运而生。动态 IP 地址是在连接因特网的时候才分配一个 IP 地址，一台计算机在多次连接因特网的时候获取的动态 IP 地址是不同的。静态 IP 地址分配给一台计算机后，无论该计算机连接因特网与否，其静态 IP 地址都不能分配给其他计算机使用。

IP 地址的分配由 NIC（Network Information Center）负责，其中 InterNIC 负责美国及其他地区，ENIC 负责欧洲地区，APNIC 负责亚太地区（其总部在日本东京大学）。我国的 IP 地址分配机构是中国互联网信息中心（CNNIC），是 APNIC 认定的中国大陆地区唯一的国家互联网注册机构（NIR）。

现在的主流 IP 地址分为 IPv4 和 IPv6 两种。IPv4 地址是一个 32 位整数，其地址格式是“W.X.Y.Z”，其中 W～Z 是一个范围为 0～255 的整数。

2011 年 2 月，全球 43 亿个 IPv4 地址资源分配完毕。这意味着因特网发展晚的国家将面临没有 IP 地址可用的问题，而且在因特网发展早期，欧、美和日本等国家分配了大量的 IPv4 地址，导致地址分配不均。为了更好地解决这个问题，人们提出了 IPv6 地址的概念。IPv6 地

[5] 数据来源于第 44 次《中国互联网络发展状况统计告》，http://www.cac.gov.cn/2019-08/30/c_1124938750.htm。

址是一个 128 位整数，其地址格式是“S:T:U:V:W:X:Y:Z”，其中 S～Z 是一个 4 位的十六进制整数。可分配的地址数量是 3.4×10^{38}，意味着每个地球人可拥有的地址数量是 5×10^{28}，从根本上解决了 IP 地址不够用的问题。

IP 地址是逻辑地址，MAC（Media Access Control）地址是物理地址，即网卡地址，是一个 48 位整数。每个网卡的 MAC 地址是全球唯一的。因特网中的任意两台计算机通信时，使用 IP 地址路由，使用 MAC 地址在同一线路上两个节点间进行通信。

任何能连接因特网的一台计算机均有一个 IP 地址和一个 MAC 地址，MAC 地址不变，除非更换网卡，IP 地址若是动态的，则每次连入因特网均是新的 IP 地址。查看这两个地址的方法如下。在 Windows 操作系统中，选择“开始|所有程序|附件|命令提示符”菜单命令，或者选择“开始”菜单，然后在“搜索程序和文件”中输入“cmd”后回车，在弹出的命令提示符窗口中输入“ipconfig /all”后回车，在显示的信息中找到物理地址和 IP 地址，见图 1.1。

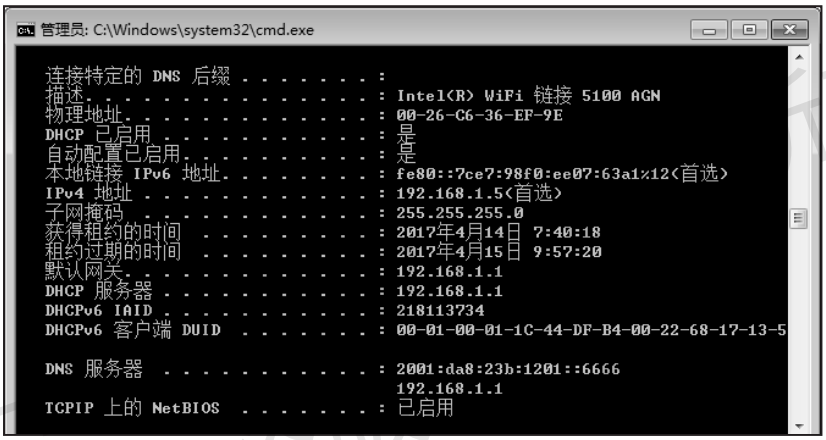

图 1.1　Windows 系统查看 IP 地址和 MAC 地址

说明：命令“ipconfig”的功能是调试计算机网络，通常用于显示计算机中网络适配器的 IP 地址、子网掩码及默认网关。这是命令不带参数的用法。

命令 ipconfig 不带任何参数选项使用时，仅显示 IP 地址、子网掩码和默认网关。如果带“all”参数，则显示完整的 TCP/IP 配置信息，除了上述信息，还包括 IP 是否动态分配、网卡的 MAC 地址等。注意，参数与“ipconfig”命令之间使用“/”（或者“-”）隔开。

Mac 操作系统中，在终端输入“ifconfig”来显示地址信息，见图 1.2。

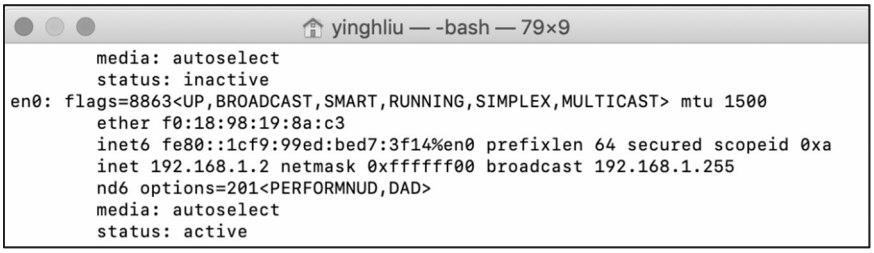

图 1.2　Mac 系统查看 IP 地址和 MAC 地址

2．协议

协议是连接在因特网上的计算机在信息交换时的统一规则和约定。协议有很多种，其中

TCP/IP（Transmission Control Protocol/Internet Protocol）是因特网最基本的协议，定义了电子设备接入因特网的方式及数据的传输标准。其中，TCP 负责传输信息，IP 负责路由。而地址解析协议（Address Resolution Protocol，ARP）用于映射 IP 地址和 MAC 地址。

1.7 域名和域名系统

IP 地址虽然能唯一定位因特网中的一台计算机，但是无论是 IPv4 还是 IPv6，一长串的数字太难记住和使用。如中华人民共和国中央人民政府门户网站（简称中国政府网）的 IP 地址是 124.202.164.208，为简化用户使用和方便记忆，域名是 www.gov.cn。注意，域名比 IP 地址要好记得多。域名（Domain Name）由一串用“.”分隔的数字或文字组成，能唯一定位因特网上一台计算机。域名系统（Domain Name System，DNS）是因特网上域名和 IP 地址一一映射的一个分布式数据库，任何人输入方便记忆的域名就能够通过域名系统转换为 IP 地址。

1983 年，保罗·莫卡派乔斯（Paul Mockapetris）发明了域名系统，但直到 1993 年随着 WWW 协议的出现，域名才开始得到各国的认可和重视。

1990 年 11 月，我国钱天白教授在国际互联网络信息中心（InterNIC）的前身 DDN－NIC 注册登记了我国的顶级域名.CN。1994 年 5 月，中国科学院计算机网络信息中心完成了中国国家顶级域名 cn 服务器的设置，结束了一直放在国外（德国 Karlsruhe 大学）的历史。

顶级域名是域名最右边的后缀名，主要分为两类。一是国际域名（iTLD，international Top-Level Domain-names，也称为国际顶级域名，是最早使用且使用最广泛的域名。国际域名按用途分类，没有国家标识。例如，com 用于商业公司，net 用于网络服务，org 用于非营利组织协会等。二是国内域名（nTLD，national Top-Level Domain name），也称为国内顶级域名，是按照国家和地区分配的后缀名。200 多个国家和地区都按照 ISO3166 国家代码分配了顶级域名。例如，中国内地的国内顶级域名是 CN。

二级域名是顶级域名左边的后缀名，中国内地的二级域名主要分为两类：一是类别域名，包括 ac（科研机构）、com（工商金融业）、edu（教育机构）、gov（政府机构）、net（互联网络信息中心和运行中心）、org（非营利组织）等；二是行政区域名，分别对应各省、自治区和直辖市，共 34 个。例如，域名 www.gov.cn 中的 cn 是顶级域名，表示中国，gov 是二级域名，表示政府机构。

1.8 网络速率

网络速率简称网速，常用的单位有 kbps、Mbps 和 Gbps。其中，bps 表示每秒钟传输的比特数量。如带宽是 1M，表示 1 Mbps，即每秒钟传输的数据量是 1 Mbits（2^{20} bits）。截至 2019 年第一季度，我国固定宽带网络平均可用下载速率为 31.43 Mbps [6]。

网络速率的测量标准并未统一，测量的网络速率也存在一定的差异。2019 年 8 月底，我国宽带发展联盟发布了第 24 期《中国宽带速率状况报告》（2019 年第二季度）。报告显示，

[6] 数据来源于第 44 次《中国互联网络发展状况统计报告》。

“2019 年第二季度，我国固定宽带网络平均下载速率达到 35.46 Mbps，固定宽带接入速率达到了 100 Mbps。两个数据值差异较大的主要原因是用户上网时的下载速率通常低于宽带接入速率值。”全球知名互联网网速测速公司 Ookla 给出了 2019 年 8 月世界网速最快的国家（或地区）名单。固定宽带第一的国家（或地区）是新加坡，平均网速是 193.90 Mbps。中国排名第 24，平均网速是 91.88 Mbps。移动网速排名第一的国家是韩国，平均网速是 111.00 Mbps。中国排名第 27，平均网速是 39.98 Mbps。

网速分为网络上行速率和网络下行速率两种。网络上行速率是指用户的计算机向因特网发送信息时的比特传输速率，如使用 FTP 工具上传文件；网络下行速率是用户的计算机下载文件比特传输速率，网速测速一般均指下行速率。一般来说，大部分网络提供商会从利润等角度考虑限制网络上行速率。

1.9 数据可视化

数据可视化是数据科学领域的一个重要分支，旨在借助图表和图形化的手段，清晰有效地传达与沟通信息。数据可视化可视化是一个复杂而漫长的过程，首先需要理解模拟信息和数字化数据等基础知识，然后掌握数据获取的技巧和方法，使用多种数据清洗方法去除“脏”数据，通过数据分析了解数据的整体特征，理解可视化基础并在符合可视化原则的基础上运用可视化工具完成数据可视化作品，最后将作品发布在网络上。

小 结

本章首先介绍了模拟和数字的概念，然后阐明了数模转换的必要性及采样和量化两个步骤，采样率和位深与文件质量和文件大小成正比。数模转换后要输入计算机处理和存储，所以我们进一步学习了进制和存储单位。计算机一般采用二进制处理数据，采用十六进制存储数据。共享计算机中的数据时，我们要掌握因特网、IP 地址、协议、域名、网速和数据可视化等基础知识。建议读者根据个人的基础和兴趣，选学地址解析协议、子网掩码等内容，了解网络包含的表面网、深网和暗网三个分层。

习 题 1

1．北京的小明给上海的丽丽发了一封邮件，分析邮件传输过程中使用到了哪些本章学习的内容？

2．小明使用数码相机拍照，照片是模拟的还是数字的？

3．小明发现某杂志的封面特别漂亮，有哪些办法可以将封面存储到计算机中？在模数转换过程中，采样率和位深是如何体现的？

4．计算机中常见的存储单位有哪些？存储单位之间如何转换？

5．IPv6 地址包含多少位整数？它比 IPv4 地址复杂又难记，为什么要使用它？