

第 3 章

计算机网络基础及数据获取

计算机网络的出现以及网络技术的普遍应用,改变了人们的生活、学习和工作方式,拓展和改变了人与人之间的沟通方式,颠覆了人们的传统思维。它已经成为现代社会人们生活和工作中必不可少的基本工具和基本技术。互联网的发展史就是一部创新史。变革从萌芽、发展再到实现,创新一直都是最直接的驱动力。而今的社会在大数据、物联网、云计算和人工智能的推动下,几年就会来一个大变样。这样一个快速发展、快速变革的时代,计算机网络的地位和作用无可替代。

本章将介绍计算机网络与 Internet 的基础知识,以及如何利用网络爬虫获取来源于互联网的海量数据信息,为后续的数据分析和处理提供基础数据。

3.1 互联时代

3.1.1 工业 1.0、2.0、3.0 时代

1769 年,1869 年,1969 年……时代的时间印记总是那么的戏剧化和序列化。这 3 个看似平淡无奇的数字,对于人类的现代文明来说,是多么的重要。它们记录了人类工业革命发展史上的伟大年代。

1769 年,瓦特制造出第一台真正意义上的蒸汽机。以蒸汽机为动力的机械生产带来了第一次工业革命,人类社会开始从手动劳动向机械生产迈进,一个崭新的工业时代在蒸汽机的隆隆巨响中开启。

整整 100 年之后的 1869 年,世界上第一条流水线在美国的辛辛那提的一家屠宰场开始使用,人类正式进入了分工明确、大批量生产的“电气时代”,电能被广泛应用在生产过程中。1866 年,德国西门子公司的第一台人类历史上的交流发电机,使得电器开始取代机器,电气动力取代蒸汽动力。加上这个流水生产线带来的劳动分工,以电气化为标志的第二次工业革命开始了,社会面貌随之发生了翻天覆地的变化。

200 年后的 1969 年,世界上第一台可编程逻辑控制器(Programmable Logic Controller,

PLC)Modicon 084 的问世,标志着人类科技文明的又一次腾飞。电子和信息技术的发明与应用,带来了产品和生产的高度自动化,迄今为止已经持续半个多世纪的第三次工业革命拉开了序幕。这次革命影响深远,将人类历史带入了史无前例的信息化时代。

这3个以一百年为跨度的特殊时期被后人总结为人们今天所熟知的“工业1.0”“工业2.0”和“工业3.0”时代。

3.1.2 ARPAnet

在1969年,还发生了一件具有划时代意义的事件。

1957年10月4日,第一颗人造地球卫星Sputnik-1在拜科努尔航天中心发射升空,差不多每天都要在美国上空飞过一次。时任美国总统的艾森豪威尔正式向美国国会提出建立国防高级研究计划局(Defense Advanced Research Projects Agency,DARPA,也常被称为ARPA)。希望通过该机构,确保不再发生毫无准备地看着他国的卫星在本国上空飞过这种让美国人无奈又尴尬的事。美国国会给了DARPA首笔520万美元的启动经费,2亿美元的项目总预算。1969年11月,全球第一个包交换网络——ARPAnet(Advanced Research Projects Agency Network,音译为阿帕网)的第一条稳定连接建立了。两周后,包含4个结点的阿帕网雏形建成。

阿帕网就是Internet的始祖。被尊称为互联网之父的文顿·瑟夫(图灵奖获得者)等人都曾参与阿帕网的研究设计。

3.1.3 TCP/IP 的出现

1974年,TCP的第一份详细说明由文顿·瑟夫和他的同事正式发表。在这份报告中,提出了“传输控制协议(Transmission Control Protocol,TCP)”和“网际协议(Internet Protocol,IP)”,即当前互联网发展的重要基石——TCP/IP。

1983年,TCP/IP正式成为Internet的标准协议,这一年,被称为Internet的元年。“万物互联”时代从此拉开了序幕。

今天,互联网在人们的生产、生活中的重大意义、深远影响不言而喻。互联网已经成为与人们的衣食住行同样不可或缺的必需品。

根据前三次工业革命那显而易见的数学规律,似乎下一次工业革命理应发生在2069年前后,这样才不会破坏数字美感。而事实上,计算机领域的金科玉律——摩尔定律,说明了事物的发展是越来越快的。摩尔定律:集成电路上可容纳的电晶体(晶体管)数目约每18个月便会增加一倍(戈登·摩尔,英特尔公司联合创始人)。

3.1.4 工业 4.0

新的转折点已来临,新一轮工业革命正悄然袭来。历史上每一次工业革命都前前后后经历了几十年的时间。所以,无论接下来的这一浪潮是否会被历史确认为第四次工业革命,人们都相信,一场深刻的技术变革正在发生,而当前正处于这场变革的开端。

关于这场变革,每个国家都赋予了它不同的说法。德国叫“工业4.0”,美国称之为“工业互联网”,而我国则提出了“中国制造2025”“从中国制造到中国智造”等。忽如一夜春风

来,千树万树梨花开。各国从政府到民众,无不在纷纷畅想和谈论未来制造业的景象。在人类历史上,还从未像今天这样信息透明到全球可以共商未来。新工业革命已经扑面而来。

而信息之所以如此透明,根本就在于互联网这一关键核心要素和关键推手,使得人类知识网络化,从而能够对这样一场未知的变革进行如此透明、透彻的探讨。

1. 德国的工业 4.0

工业 4.0 包括将信息物理系统(Cyber Physical System,CPS)技术一体化应用于制造业和物流行业,以及在工业生产过程中使用物联网和服务技术。这种一体化的充分融合,从横向上来看,能实现价值链上的企业间的集成;从纵向上来看,能实现网络化制造系统的集成。另外,端对端的工程数字化集成也有助于打造工业产品和服务全面交叉渗透的“智能工厂”和“智能生产”,从而推进生产或服务模式由集中式控制向分散式控制转变,实现高度灵活的个性化和数字化生产及服务,最终使生产更智能、更高效、更快速、更经济。

工业 4.0 的主要愿景和关键技术就是:这是一个由物联网、互联网和云计算等技术连接起来的网络化、分布式的智能生产系统。在该系统中,机器或产品具有自组织、自优化、自配置和自诊断的智能。

2. 美国的工业互联网

2000 年,一个叫沙利文(Frost & Sullivan)的美国公司开创性地提出了“工业互联网”这个概念——用以指代复杂物理机器和网络化传感器及软件的集成。2012 年 11 月 26 日,通用电气发布了《工业互联网:打破智慧与机器的边界》白皮书,正式提出了“工业互联网”的概念。2014 年,通用电气联合 AT&T、Cisco、Intel 和 IBM 等知名企业在美国波士顿成立了工业互联网联盟(Industrial Internet Consortium,IIC)。

3. 中国制造 2025

2015 年 5 月,国务院正式发布《中国制造 2025》规划,作为我国工业未来 10 年的发展纲领和顶层设计,目的就在于将我国从制造大国转型为“制造强国”。

“中国制造 2025”,一个把时间、地点、事件三要素都交代得清清楚楚的命名,充分结合了我国的国情,立足于我国的现状,其目标更明确,内涵更确切,路线也更清晰。该战略规划“以促进制造业创新发展为主题,以提质增效为中心,以加快新一代信息技术与制造业深度融合为主线,以推进智能制造为主攻方向,以满足经济社会发展和国防建设对重大技术装备的需求为目标,强化工业基础能力,提高综合集成水平,完善多层次多类型人才培养体系,促进产业转型升级,培育有中国特色的制造文化,实现制造业由大变强的历史跨越”。这一规划不仅包括我国实施制造强国战略的指导思想、基本方针,还指出了重点领域和机遇及挑战。

3.2 计算机网络概述

最初计算机网络的诞生,是为了实现资源共享。而随着计算机技术和网络技术的发展,今天,人们通过计算机网络不仅可以共享硬件资源、软件资源,而且可以共享大量的信息资源。特别是随着移动网络和无线网络的应用和普及,计算机网络渗透到了人们的生活、学习以及社会的方方面面。可以说,现在人们如果离开了网络,就会寸步难行。

3.2.1 计算机网络的定义

计算机网络的精确定义并未统一。

1. 计算机网络的最简单的定义

计算机网络是一些互相连接的、自治的计算机的集合。

“自治”指的是具有完整的软硬件系统,可以单独运行、使用的独立的计算机;“互相连接”指的是计算机之间能进行数据通信和信息交换。

2. 更为准确的计算机网络的定义

将分布在不同位置的、具有独立功能的计算机,通过通信设备和通信线路连接起来,完成信息交换,以实现资源共享的协同工作的计算机集合。

更具体地说,计算机网络是由若干结点(Node)通过网络适配器(简称网卡),利用各种不同材质的作为传输介质的通信线路,以各种形式的拓扑结构连接而成的网络。计算机网络中的结点可以是计算机、集线器、交换机和路由器等。

3.2.2 计算机网络的功能与作用

1. 主要功能

计算机网络面向用户提供的主要功能有两个:

(1) 连通性: 计算机网络使上网用户之间可以交换信息,好像这些用户的计算机都可以彼此直接连通一样。用户之间的距离也因此而变得更近了。

(2) 共享: 即资源共享。共享的含义是多方面的。可以是信息共享、软件共享,也可以是硬件共享。由于网络的存在,这些资源好像就在用户身边一样。

2. 其他功能

1) 提高系统的处理能力

计算机网络的出现,使得以前单个计算机无法处理和完成的事情,现在可以利用网络中的若干台机器来共同处理和完成,从而提高了系统的处理能力。

2) 提高系统的可靠性

在没有计算机网络之前,在一个系统内,在单个部件或者计算机暂时失效时,必须通过更换资源的办法来维持系统的继续正常运行。而在计算机网络中,各种资源(尤其是程序与数据)可以存放在多个地点,用户可以通过多种途径来访问网内的某个资源。当计算机网络中的某台设备出现故障时,不会影响整个网络的运行,借助冗余和备份就可以提高系统的可靠性。

3.2.3 数据通信

数据通信是通信技术与计算机技术相结合而产生的通信方式。数据通信是指在两个计算机或终端之间以二进制数的形式进行信息交换、传输数据。下面介绍几个关于数据通信的常用术语。

1. 信道及信道的通信方式

信道(Channel)和电路并不等同。信道一般都是用来表示向某一个方向传送信息的媒

体。因此,一条通信电路往往包含一条发送信道和一条接收信道。

从通信双方信息交互的方式来看,有以下 3 种基本方式:

(1) 单向信道:又称为单工通信,即只能有一个方向的通信而没有反方向的交互。无线电广播或有线电广播及电视广播就属于这种类型。

(2) 双向交替信道:又称为半双工通信,即通信的双方都可以发送信息,但不能双方同时发送(当然也不能同时接收)。这种通信方式是一方发送另一方接收,过一段时间再反过来。

(3) 双向同时通信:又称为全双工通信,即通信的双方可以同时发送和接收信息。

2. 带宽与数据传输速率

带宽指的是信道的最高频率和最低频率之差,即频率的范围,其基本单位是赫兹(Hz)。信道的带宽越宽,传输的数据量就越大。

在数字信道中,用数据传输速率(比特率)表示信道的传输能力,即每秒传输的二进制位数(b/s)。香农定理证明,最大数据传输速率与信道带宽之间存在着明确的正比关系,所以人们也经常用“带宽”来表示信道的数据传输速率。现在所说的带宽通常指的就是数据传输速率。

3.2.4 计算机网络的分类

计算机网络的分类方法有很多,下面简单介绍几种常见的分类。

1. 按网络的作用范围进行分类

(1) 广域网(Wide Area Network, WAN)。广域网的跨接很大,作用范围通常为几十千米到几千千米,因而有时也称为远程网。广域网是 Internet 的核心部分,其任务是通过长距离(例如,跨越不同的国家)运送主机所发送的数据。连接广域网各结点交换机的链路一般都是高速链路,具有较大的通信容量。

(2) 城域网(Metropolitan Area Network, MAN)。城域网的作用范围一般是一个城市,可跨越几个街区甚至整个城市,其作用距离为 5~50km。城域网可以为一个或几个单位所拥有,也可以是一种公用设施,用来将多个局域网进行互连。目前,很多城域网采用的都是局域网技术,因此城域网有时也常纳入局域网的范围进行讨论。

(3) 局域网(Local Area Network, LAN)。局域网一般用微型计算机或者工作站通过高速通信线路相连(现在速率通常都在 100Mb/s 以上),但地理范围上在几百米到十几千米内。局域网一般具有高数据传输率、低延迟和低误码率的特点。

2. 按网络的使用者进行分类

(1) 公用网(Public Network)。这是指对全社会开放并提供服务的网络,如国家电信部门出资建造的大型网络。公用的意思就是所有愿意按规定缴纳费用的人都可以使用这种网络。因此公用网也称为公众网,如 CHINANET。

(2) 专用网(Private Network)。这是某个部门、某个行业为各自的特殊业务工作需要而建造的网络。这种网络不对外人提供服务,如军队、政府、银行、铁路、电力、公安等系统都有本系统的专用网。

3. 按通信方式进行分类

(1) 客户/服务器(Client/Server,C/S)方式。这种方式在 Internet 上是最常用的,也是最传统的方式。在网上发送电子邮件或者在网站上查找资料的时候,都是使用客户/服务器方式。客户是服务请求方,服务器是服务提供方。

(2) 对等连接方式(Peer-to-Peer,P2P)。对等连接方式是指两个主机在通信时并不区分哪一个是服务请求方,哪一个是服务提供方。只要两个主机都运行了对等连接软件(P2P 软件),它们就可以进行平等的对等连接通信。

3.3 Internet 基础

3.3.1 IP 地址

【情景故事 3-1】 小红刚搬了新家,安顿完所有家当后,小红遇到了难题——没有网络。对于互联网几乎一窍不通的她决定向她的“网络大神朋友”小明请教,于是有了下面的对话:

小红:小明,你好,我刚搬了新家,但是还没有联网,你比较在行,想问问你咋办啊?

小明:哦,你是要连接 Internet 啊,那你首先需要连接到某一个 ISP,来获得上网所需的 IP 地址。

小红:ISP 是什么?还有 IP 地址,老听到这个词,但就是不知道到底是什么意思。

小明:ISP 是“互联网服务提供商”,也就是我们熟悉的“中国电信”“中国移动”和“中国联通”,等等,我们要上网首先需要向它们提出申请,由它们为我们提供网络的接入和信息业务服务,这样我们就能使用网络功能啦。IP 地址就比较重要了,它是给每个连接在 Internet 上的主机(或路由器)分配的一个在全世界范围内唯一的标识符。比如我在 QQ 上给你发消息,消息中的信息是如何传送到你的计算机中的呢?要在连接互联网的亿万台主机中找到你那一台计算机,就需要知道代表你那台计算机的唯一标识符,这就是 IP 地址。这就好比我要给你打电话,就必须知道你的电话号码一样。IP 地址就是每台主机在 Internet 中的电话号码,有了它你就能收发信息了。

小红:原来是这样,IP 地址就像电话号码一样,电话号码是由数字组成的,而且固定电话一般分为两个部分,前面是区号,后面才是实际的电话号码,IP 地址也是这样表示的吗?

小明:嗯嗯,它们还真有些类似呢。下面我们就来看看 IP 地址的表示和编码方案吧。

1. IP 地址的表示

同电话号码一样,IP 地址使用固定长度的数字来表示,现在常用的表示方法为 IPv4 地址,它是一个由 32 位(b)二进制数组成的地址。人们在实际应用中为了便于表达,一般将这 32b 数分为 4 段,每段 8b,然后将这 4 段 8b 的二进制数转换为十进制数,十进制数之间用“.”区隔。这种方法叫作点分十进制表示法(Dotted Decimal Notation)。例如,地址 10000000 01100100 00000011 00001010 用点分十进制数表示为 128.110.3.10。

IP 地址采用层次结构,由两部分构成,即网络号与主机号,网络号在前,主机号在后。其中,网络号用来标识主机所在的逻辑网络(类似于固定电话号码前的区号),主机号用来表

示网络中的一个接口。一台 Internet 主机至少有一个 IP 地址,而且该 IP 地址是全球唯一的。如果一台 Internet 主机有两个或多个 IP 地址,则该主机属于两个或多个逻辑网络。

2. IP 地址的编码方案

传统的 IP 地址编码方案采用所谓的“分类 IP 地址”,分别称为 A 类、B 类、C 类、D 类和 E 类。其中 A 类、B 类和 C 类由全球性的地址管理组织在全球范围内统一分配,D 类和 E 类属于特殊地址。

IP 地址采用高位字节的高位来标识地址类别。IP 地址分类编码方案如图 3-1 所示。

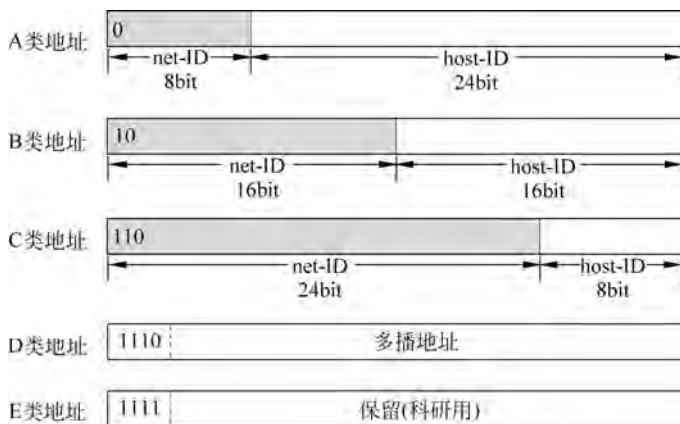


图 3-1 IP 地址分类编码方案

(1) A 类地址的第一位为 0,B 类地址的前 2 位为 10,C 类地址的前 3 位为 110,D 类地址的前 4 位为 1110,E 类地址的前 4 位为 1111。

(2) A 类、B 类和 C 类地址的网络号字段分别为 1 字节(B)、2 字节和 3 字节,A 类、B 类和 C 类地址的主机号字段分别为 3 字节、2 字节和 1 字节。

将 IP 地址划分为 3 个类别的原因是:各种网络的差异很大,有的网络拥有很多主机,而有的网络上的主机很少,将 IP 地址划分为 A 类、B 类和 C 类可以更好地满足不同用户的要求。

当某个单位申请到一个 IP 地址时,实际上只是获得了一个网络号(net-ID),具体的各个主机号(host-ID)则由该单位自行分配,只要做到在该单位范围内无重复的主机号即可。

除了上述 3 类 IP 地址以外,还有两类使用的较少的地址,即 D 类和 E 类地址。D 类地址是多播地址,E 类地址保留给以后使用。

A 类地址的 net-ID 字段有 1 字节,由于最高位已经固定为 0,因此剩下的 7 位共能表示 $126(2^7 - 2)$ 个 A 类网络,这里减 2 的原因是:全 0 的 IP 地址是保留地址,意思是“本网络”;值为 127(即 01111111)的地址保留作为本地软件环回测试(Loopback Test)本主机之用。后 3 个字节是 host-ID,每一个 A 类网络中的最大主机数量是 $16\,777\,214(2^{24} - 2)$ 。减 2 的原因是:全 0 的 host-ID 字段表示该 IP 地址是“本主机”所连接到的单个网络地址(例如,某一主机的 IP 地址是 126.100.10.8,则该主机所在的网络地址就是 126.0.0.0),而 host-ID 为全 1 表示“所有的(All)”,因此全 1 的 host-ID 字段表示该网络上的所有主机,即本网内广播。整个 A 类地址空间共有 2^{31} (即 2 147 483 648)个地址,而 IP 地址全部的地址空间共有 2^{32} (即 4 294 967 296)个地址。可见 A 类地址占有整个 IP 地址空间的 50%。

B类地址的 net-ID 字段有 2 字节,但前面 2b 值已经固定(10),只剩下 14b 可以变化,因此 B 类地址的网络数为 $16\,384(2^{14})$ 。注意,这里不需要减 2,因为这 14b 加上最前面固定的 2b 值 10,无论如何也构不成全 0 或者全 1。B 类地址的每一个网络上的最大主机数是 $65\,534(即\ 2^{16}-2)$ 。这里减 2 和 A 类网络一样是因为要扣除全 0 和全 1 的主机号。整个 B 类地址空间共有 $1\,073\,741\,824(2^{30})$ 个地址,占整个 IP 地址空间的 25%。

C 类地址有 3 字节的 net-ID 字段,最前面 3b 的标识位是 110,还有 21b 可以变化,因此 C 类地址的网络总数是 $2\,097\,152(即\ 2^{21})$,这里也不需要减 2)。每一个 C 类地址的最大主机数是 $254(即\ 2^8-2)$ 。整个 C 类地址空间共有 $536\,870\,912(即\ 2^{29})$ 个地址,占整个地址空间的 12.5%。

所有 IP 地址的使用范围如表 3-1 所示。

表 3-1 所有 IP 地址的使用范围

网络类别	最大网络数	第一个可用的网络号	最后一个可用的网络号	每个网络中的最大主机数
A	$126(2^7-2)$	1	126	16 777 214
B	$16\,384(2^{14})$	128.0	191.255	65 534
C	$2\,097\,152(2^{21})$	192.0.0	223.255.255	254

一般不使用的特殊 IP 地址如表 3-2 所示。

表 3-2 一般不使用的特殊 IP 地址

net-ID	host-ID	源地址使用	目的地址使用	代表的意义
0	0	可以	不可以	在本网络上的本主机
0	host-ID	可以	不可以	在本网络上的某个主机
全 1	全 1	不可以	可以	只在本网络上进行广播(各路由器均不转发)
net-ID	全 1	不可以	可以	对 net-ID 上的所有主机进行广播
127	任何数	可以	可以	用作本地软件环回测试

【情景故事 3-2】 小红:哇!那一旦我的计算机被分配到一个 IP 地址,这个地址就永远不变,一直属于我,并且永远可以定位出我的主机了吗?

小明:IP 地址当然不是永远不变的,它描述的是定位某一网络主机的方式,所以当主机的地理位置发生改变时,IP 地址就可能变化。即使主机的地理位置保持不变,但如果改变了连接的线路,即连接到另外一个网络,那么这个主机的 IP 地址也必须改变。而且,现在连接到 Internet 上的设备数量越来越多,IP 地址已经不足以分配每一台主机了,现在咱们上网一般分配的是动态 IP 地址,每次上网使用的 IP 地址都是临时的,会随时间变化。

小红:那么多 IP 地址都不够用了吗?真是难以想象。还有动态 IP 是什么原理呢?能具体讲讲吗?

小明:好的,看你这么感兴趣,下面我们就来详细说说 IP 地址危机和它的解决方法。

3. IP 地址危机

随着 IP 网络爆炸性的发展,更重要的是全球 Internet 的飞速发展,可用的 IP 地址空间正在缩小,核心的 Internet 路由器处理能力也逐渐耗尽。Internet 面临着必须尽早解决的问题,这就是:

(1) IPv4 网络地址的耗尽问题。

(2) 由于 Internet 的发展,Internet 的路由选择表的大小在迅速、大量地增加。随着更多的 C 类地址加入 Internet,新网络信息的大量充斥威胁到 Internet 路由器的处理能力。

在 IPv4 地址结构下,A 类和 B 类地址构成了 75% 的 IPv4 地址空间,但只有少数公司和组织能够分配到一个 A 类或 B 类网络号。C 类网络号比 A 类和 B 类网络号要多得多,但它们仅仅占了可能的 40 亿 (2^{32}) 个 IP 地址的 12.5%,各类地址所占比例如图 3-2 所示。

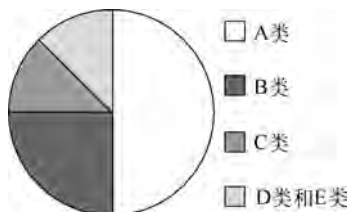


图 3-2 各类地址所占比例

2019 年 11 月 25 日,RIPE NCC 宣布,其最后的 IPv4 地址空间储备池已完全耗尽,所有 43 亿个 IPv4 地址已分配完毕。

人们一直在寻求解决 IPv4 地址危机的办法,常用的方法有:

- ① 无类域间路由(CIDR)和可变长子网掩码(VLSM)。
- ② 私有 IP 地址^①与网络地址转换(NAT)。

人们在 A 类、B 类和 C 类地址段中各取了一部分地址空间作为私有地址。这部分规划保留的地址是:

A 类 IP 地址中的 10.0.0.0~10.255.255.255;

B 类 IP 地址中的 172.16.0.0~172.31.255.255;

C 类 IP 地址中的 192.168.0.0~192.168.255.255。

私有地址不能直接接入 Internet,也不会被 Internet 路由。使用了私有 IP 地址的本地网络中的计算机如果需要连接 Internet,需要借助于专门的技术,即网络地址转换技术(Network Address Translation,NAT)。

NAT 允许一个整体的本地网络在其内部均使用私有 IP 地址,在 Internet 上只使用一个或少量的公用 IP 地址。当内部结点需要与外部网络进行通信时,NAT 可将内部私有 IP 地址翻译成外部公有 IP 地址,从而得以正常访问 Internet。这样一来,就可以使用较少的公有 IP 地址,解决更多内部结点机器的 Internet 访问问题,从而有效地缓解了 IP 地址不足的问题。

- ③ IPv6(彻底的根本解决方法)。

IPv6 把原来的 IPv4 地址增大到了 128 位(b),其地址空间大约是 3.4×10^{38} ,是原来 IPv4 地址空间的 2^{96} 倍,这样就可以彻底解决地址不足的问题。另外,IPv6 并没有完全抛弃原来的 IPv4,并且在若干年内都会与 IPv4 共存。IPv6 使用一系列固定格式的扩展首部取代了 IPv4 中可变长度的选项字段。IPv6 对 IP 数据报协议单元的头部进行了简化,仅仅包含 7 个字段(IPv4 有 13 个)。这样,当数据报文经过中间的各个路由器时,各个路由器对其处理的速度可以更快,从而可以提高网络吞吐率。IPv6 内置了支持安全选项的扩展功能,如身份验证、数据完整性和数据机密性等。

^① 所谓私有地址就是在 A、B、C 3 类 IP 地址中保留下来为内部网络分配地址时所使用的 IP 地址。私有地址主要用于在局域网中进行分配,在 Internet 上是无效的。这样可以很好地隔离局域网和 Internet。私有地址在公网上是不能被识别的,必须通过 NAT 将内部 IP 地址转换为公网上可用的 IP 地址,从而实现内部 IP 地址与外部公网的通信。

4. 静态 IP、动态 IP 和 DHCP

将一台计算机连接到 Internet, 不管它是哪种机型, 也不管它是通过什么连接方式与 Internet 相连, 都必须首先配置一个 IP 地址。对于所制定并配置的 IP 地址, 可以有两种选择: 静态 IP 地址配置或动态 IP 地址配置。

静态 IP 地址配置就是将一个 IP 地址“永久地”“固定地”分配给某一台主机, 其他的主机不再可能使用该 IP 地址。此种情况通常出现在为多路用户访问的网络设备配置 IP 地址时。例如, 网络服务器或网络打印机等。此类设备不宜经常变更 IP 地址, 避免用户每次访问它们时需要频繁地更换 IP 地址。

静态 IP 地址配置情况下, IP 地址不能重复用。另外, 网络管理员应该妥善记录和保存设备与其配置的 IP 地址的对应关系, 以便于网络维护。

动态 IP 地址配置就是按照一定的规则, 将可分配使用的全部或部分的 IP 地址集中管理起来, 当网络中的某主机需要连接和访问其他设备进行数据通信时, 由网络管理系统自动地分配给该主机(计算机、打印机等)一个 IP 地址。这种 IP 地址的分配与使用不是固定的。当某主机中断网络连接时, 可以收回已分配的 IP, 进行重新地址分配; 而且某台主机每次网络连接所拥有和使用的 IP 地址可能不同。

动态 IP 地址配置的最大特点就是提高了有限个 IP 地址的使用效率, 并减少了网管人员的维护工作量。

在指定网络中主机 IP 地址的配置方案时, 除了对必须采用静态 IP 地址的设备配置静态 IP 外, 一般都应采用动态 IP 地址的配置方式。

动态 IP 地址配置通常是由 DHCP 服务器分配。它可以是网络中的一台计算机或其他一些硬件设备, 例如路由器。动态主机配置协议(Dynamic Host Configuration Protocol, DHCP)提供了一种自动分配 IP 地址、子网掩码、默认网关等地址信息以及其他配置信息的服务。在需要 IP 地址动态配置的方式中, 网络中至少有一台服务器安装了 DHCP 服务。

DHCP 是一个客户/服务器协议。需要服务时, DHCP 客户端将首先向 DHCP 服务器发出服务请求, 申请一个 IP 地址; 服务器接收到该请求后, 将按照预先制定好的方式, 进行回应并最终向客户端分配和提供一个 IP 地址。该客户端主机在使用完毕、即将退出网络连接时, 将再次联系 DHCP 服务器, 交还 IP 地址。

不管是动态 IP 地址配置还是静态 IP 地址配置, IP 地址配置是否正确都将影响到网络的运行。IP 地址的配置除了要避免重复之外, 还应遵守以下规则:

- 每个网络的网络号必须是唯一的。
- 网络号不能为全 0(网络号为全 0 表示一个本地网)。
- 网络号不能为全 1。
- 网络号不能以 127 开头(127 是环回地址)。
- 一个网络中, 每台主机的主机号必须是唯一的。
- 主机号不能全为 1(主机号全为 1 表示广播地址)。
- 主机号不能全为 0(主机号全为 0 表示网络地址)。

【情景故事 3-3】 小红: 我刚才看到了一个叫子网掩码的词, 它经常和 IP 地址放在一起说, 这是为什么呢? 它和 IP 地址又有什么关联呢?

小明：好问题。我们刚才已经知道 IP 地址是分类管理的，不同类型的 IP 地址属于不同规模的网络。一个 A 类网络下有 1600 多万个主机，如此庞大的网络管理起来是十分困难的，所以一般会在网络下继续划分子网，并将 IP 地址主机号中的前几位作为子网号进行标识。这样，每个 IP 地址就由三个部分组成，前面是网络号，中间是子网号，最后是主机号。那么，子网号究竟在主机号中占前几位呢，直接通过 IP 地址是看不出来的。而子网掩码就是用来确定子网号在 IP 地址中对应位置的。

小红：原来如此，那么子网掩码是怎么表示的？如何用子网掩码确定网络号和子网号呢？

小明：别急，下面我们就来看看子网掩码的详细内容。

5. 子网掩码

子网掩码和 IP 地址的表示方法类似，同样由 32 位二进制数组成。子网掩码中的 1 对应于 IP 地址中的网络号和子网号，而子网掩码中的 0 对应于 IP 地址中的主机号。注意，虽然在互联网标准文档(RFC)中并没有规定子网掩码必须是由连续的 1 组成的，但是在实际应用中一般选择连续的 1 以免出现可能发生的差错。

网络地址(在划分子网的情况下常称为子网地址)就是将主机号 host-ID 置为 0 的 IP 地址。这也就是将子网掩码和 IP 地址按位相“与”(AND)的结果(对于二进制数而言，任何数与 0 相“与”结果为 0，任何数与 1 相“与”结果为其本身)。比如一个 IP 地址为 138.114.241.1 (10001010 01110001 11110001 00000001)，子网掩码为 255.255.248.0 (11111111 11111111 11111000 00000000)，则通过按位“与”得到该 IP 地址对应的网络号为 138.114.240.0 (11001010 01110001 11110000 00000000)。通过子网掩码获取网络号过程如图 3-3 所示。

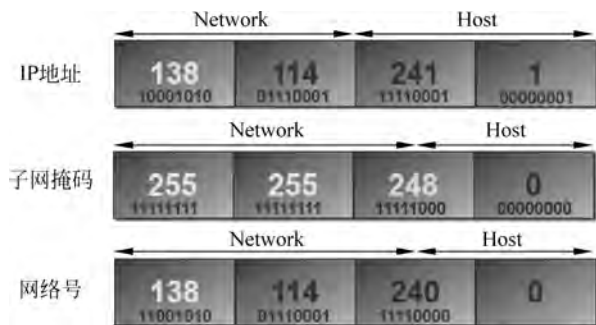


图 3-3 通过子网掩码获取网络号过程

那么，为什么在没有划分子网的时候还要使用子网掩码呢？这样做的目的是简化路由器的路由选择算法。因此，Internet 的标准规定，所有的网络都必须有一个子网掩码，同时在路由器的路由选择表中也必须要有子网掩码这一栏。如果一个网络不划分子网，那么该网络的子网掩码就使用默认子网掩码。默认子网掩码中的 1b 的位置和 IP 地址中的网络号字段正好一一对应。因此，若将默认的子网掩码和某个不划分子网的 IP 地址按位“与”，就得出该 IP 地址的网络地址来。这样做可以不用查找该地址的类别就能知道这是哪一类的 IP 地址。显然，A 类地址的默认子网掩码是 255.0.0.0，B 类地址的默认子网掩码是 255.255.0.0，C 类地址的默认子网掩码是 255.255.255.0。

子网掩码是一个网络或一个子网的重要属性。路由器在和相邻路由器交换路由信息时，必须将自己所在网络(或子网)的子网掩码告诉相邻路由器。在路由器的路由选择表中

的每一个项目,除了要给出目的网络地址外,还必须同时给出该网络的子网掩码。若一个路由器连接在两个子网上就拥有两个网络地址和两个子网掩码。

【情景故事 3-4】 小红:我明白了,子网掩码就像一个罩子一样,和 IP 地址搭配在一起,盖住主机号,只露出网络号?

小明:可以这么理解,不论网络是否划分了子网,不论网络号长度是 1 字节、2 字节还是 3 字节,只要将子网掩码和 IP 地址进行按位“与”运算,就会立即得出网络地址来。

小红:对了,我还听说除了 IP 地址之外,还有一种地址叫作 MAC 地址?

小明:没错,MAC 地址是和网络物理设备相关的地址,关于它的解释如下。

6. MAC 地址

MAC 地址又称为硬件地址或者物理地址。它是一个 48 位的全球地址,是指局域网上的每一台计算机中固化在适配器(网卡)的 ROM 中的地址。更准确地说,这种 48 位地址应当是某个接口的标识符。

这个 48 位地址中的前 3 字节的 24 位是由 IEEE 的注册管理机构 RA 负责分配的,也就是全世界生产网卡的厂家都必须向 IEEE 购买由这 3 字节构成的这个号。地址字段中的后 3 字节则是由厂家自行指派。

全世界所有网卡的 MAC 地址都是不同的。也就是说,没有完全相同的两块网卡,这也是在网络中能够准确地找到主机的原因。

【情景故事 3-5】 小红:我大概明白了,MAC 地址也是一种标识符,它是和计算机硬件也就是网卡绑定在一起的。那一台计算机的 MAC 地址和 IP 地址又有什么关联呢?

小明:IP 地址标识的是一台主机在网络中的逻辑位置,而 MAC 地址则标识主机本身。一般 IP 地址和 MAC 地址是对应的,但这种对应关系可以发生改变。就好比一个人,IP 地址代表他的住址,MAC 地址代表他本人(身份证号),你可以通过住址找到一个人,也可以通过身份证号去搜索一个人。当然一个人的住址可以发生改变,一个住址的居住者也可能发生改变。

小红:我懂了,但是既然 MAC 地址就可以唯一地标识一台主机了,那么我们为什么还要用 IP 地址呢?

小明:还是刚才的例子,你要找一个人的时候,是不是通过地址比较方便一点呢?因为地址毕竟是有规律可言的,某省、某市、某街道非常容易逐级定位,通过地址找到人之后通过身份证号认证一下你的身份就可以了。但是如果单独通过身份证号去找一个人就很困难了,茫茫人海,不可能一个个去核对身份证吧?

小红:我明白了。

3.3.2 域名系统

【情景故事 3-6】 小红:听你说了这么多,我总算明白 IP 地址的意思了。不过如果我想找到网络上的另一台主机,就必须记住它的 IP 地址吗?网络上那么多主机,IP 地址又很长,记不住怎么办呢?

小明:这个问题大家早就考虑到了,还记得我之前把 IP 地址比作电话号码的例子吗?

你平时通过手机联系你的小伙伴,是直接输入电话号码联系的吗?

小红:当然不是,那么多号码我哪里记得住?肯定是事先都存放在手机通讯录里,直接在通讯录里面找到小伙伴的名字就行了。

小明:没错,访问网络主机也是一样的道理。当我们要定位一台主机时,往往使用的是主机的域名而非 IP 地址,这个域名就好比人名一样比较容易记忆,而通讯录就是网络中的域名服务器,用来关联每个域名与 IP 地址,使用通讯录进行联系的机制就是域名系统。

小红:这么神奇?它究竟是如何工作的呢?

小明:下面我们就来具体说说什么是域名系统吧。

1. 域名系统基础

域名系统(Domain Name System,DNS)是 Internet 使用的命名系统,用来把便于人们使用的机器名字转换为 IP 地址。域名系统其实就是名字系统。

用户与 Internet 上某个主机通信时,必须知道对方的 IP 地址。然而用户很难记住长达 32 位的二进制数主机地址。即使是点分十进制数表示的 IP 地址也并不太容易记忆。但在应用层为了方便用户记忆各种网络应用,更多的是使用主机名。那为什么机器在处理 IP 数据报时要使用 IP 地址而不使用域名呢?这是因为 IP 地址的长度是固定的 32 位(IPv6 地址是 128 位),而域名的长度是不固定的,机器处理起来比较困难。

Internet 的域名系统 DNS 被设计成为一个联机分布式数据库系统,并采用客户/服务器方式。DNS 大多数名字都在本地进行解析,仅少量解析需要在 Internet 上通信,因此 DNS 系统的效率很高。由于 DNS 是分布式系统,即使单个计算机出了故障,也不会妨碍整个 DNS 的运行。

域名到 IP 地址的解析过程的要点如下:当某一个应用进程需要把主机名解析为 IP 地址时,该应用进程就调用解析程序,并成为 DNS 的一个客户,把待解析域名放在 DNS 请求报文中,以用户数据报协议(User Datagram Protocol,UDP)方式发给本地域名服务器(使用 UDP 是为了减少开销)。本地域名服务器在查找域名后,把对应的 IP 地址放在回答报文中返回。应用进程获得目的主机的 IP 地址后即可进行通信。

若本地域名服务器不能回答该请求,则此域名服务器就暂时成为 DNS 中的另一个客户,并向其他域名服务器发出查询请求。这种过程直到能够回答该请求的域名服务器为止。

【情景故事 3-7】 小红:我懂了,用域名代替 IP 地址方便大家记忆,域名就像每台主机的名字一样。那名字可以自己随便起吗?我把我的主机域名改成“小红家的主机”可不可以?

小明:这个肯定是不行的。其实人名也是不能随便起的,域名更是如此。一般 Internet 中的域名采用层次结构进行命名,下面我们就来了解一下具体的命名规则吧。

2. Internet 的域名结构

Internet 采用层次树状结构的命名方法,任何一个连接在 Internet 上的主机或者路由器,都有一个唯一的层次机构的名字,即域名。这里,“域”指的是名字空间中一个可被管理的划分。域还可以划分为子域,而子域还可继续划分为子域的子域,这样就形成了顶级域、二级域、三级域等。多级多层次域名结构如图 3-4 所示。

从语法上讲,每一个域名都是由标号序列组成,而各标号之间用“.”隔开。例如图 3-4 所示的域名就是中央电视台的邮件服务器的域名,它由 3 个标号组成,其中标号 com 是顶级域名,标号 cctv 是二级域名,标号 mail 是三级域名。



图 3-4 多级多层次域名结构

DNS 规定,域名中的标号都由英文字母和数字组成,每一个标号不超过 63 个字母(但为了记忆方便,最好不要超过 12 个字母),也不区分大小写。标号中除了连字符(-)外不能使用其他的标点符号。级别最低的域名写在最左边,而级别最高的顶级域名则写在最右边。由多个标号组成的完整域名总共不能超过 255 个字符。

DNS 既不规定一个域名需要包含多少个下级域名,也不规定每一级的域名代表什么意思。各级域名由其上一级的域名管理机构管理,而最高的顶级域名则由 ICANN(the Internet Corporation for Assigned Names and Numbers,互联网名称与数字地址分配机构)进行管理。用这种方法可使每一个域名在整个 Internet 范围内是唯一的,并且也容易设计出一种查找域名的机制。

需要特别注意的是,域名只是个逻辑概念,并不代表计算机所在的物理地点。变长的域名和使用有助记忆的字符串,是为了便于人来使用。而 IP 地址的定长的数字则非常有利于机器进行处理。

3. 顶级域名

原来的顶级域名共分为 3 大类。

(1) 国家顶级域名 nTLD。采用 ISO 3166 规定。如:cn 表示中国,us 表示美国,uk 表示英国,等等。国家顶级域名又常记为 ccTLD(cc 代表国家代码)。

(2) 通用顶级域名 gTLD。最先确定的通用顶级域名有 7 个,即 com(公司企业)、net(网络服务机构)、org(非营利性组织)、int(国际组织)、edu(美国专用的教育机构)、gov(美国的政府部门)和 mil(美国的军事部门)。

截止到 2011 年初,又陆续增加了 13 个通用顶级域名:aero(航空运输企业)、asia(亚太地区)、biz(公司和企业)、cat(使用加泰隆人的语言和文化团体)、coop(合作团体)、info(各种情况)、jobs(人力资源管理者)、mobi(移动产品与服务的用户和提供者)、museum(博物馆)、name(个人)、pro(有证书的专业人员)、tel(Telnic 公司)和 travel(旅游业)。

(3) 基础结构域名(infrastructure domain)。这种顶级域名只有一个,即 arpa,用于反向域名解析,因此又称为反向域名。

值得特别注意的是,2011 年 6 月 20 日在新加坡会议上正式批准新顶级域名(New gTLD),因此任何公司、机构都有权向 ICANN 申请新的顶级域名。新顶级域名使企业域名具有了显著的、强烈的标志特征。因此,新顶级域名被认为是真正的企业网络商标。新顶级域名是企业品牌战略发展的重要内容,其申请费用很高(约 18 万美元)。新顶级域名已经与 2013 年开始启用。

在国家顶级域名下注册的二级域名均由该国家自行规定。例如,顶级域名为 jp 的日本,将其教育和企业机构的二级域名定义为 ac 和 co,而不是 edu 和 com。

4. 我国的二级域名

我国把二级域名划分为“类别域名”和“行政区域名”两大类。

类别域名共 7 个,分别为 ac(科研机构)、com(工、商、金融等企业)、edu(教育机构)、gov(政府机构)、mil(国防机构)、net(提供互联网络服务的机构)和 org(非营利性组织)。

行政区域名一共 34 个,适用于我国的各省、自治区和直辖市。

我国修订的域名体系允许直接在 cn 的顶级域名下注册二级域名。这显然给我国的 Internet 用户提供了极大的方便。关于我国的互联网络发展现状以及各种规定(包括申请域名的手续),均可在中国互联网络信息中心 CNNIC 的网址上找到。

【情景故事 3-8】 小红:我明白了,域名真是方便呢。话说有了域名,我都不想再用 IP 地址了。

小明:不能这么说。域名对于我们虽然很方便,但对于计算机来说使用起来就很麻烦了,毕竟计算机里最容易表达的还是以 32 位二进制数组成的 IP 地址。因此在我们进行网络访问的时候,虽然使用的是域名,但最终还是需要把它转换为 IP 地址,就像你用手机通讯录联系其他人时,最终还是使用通讯录中存放的电话号码进行拨号的。

小红:手机通讯录是存在手机里的,你这里说的网络通信录也是存在我自己的主机中吗?我可以自己查看吗?

小明:这个和手机不太一样,Internet 中的通讯录叫作域名服务器,是网络中专门存放 IP 地址和域名对应关系的主机,你每次访问网络时实际需要先根据所处的位置访问本地域名服务器才能把域名转换为 IP 地址。

小红:那我要访问这个域名服务器,是不是先要知道它的 IP 地址啊?

小明:没错,这个 IP 地址也叫作 DNS 服务器地址,需要你在接入网络时预先设置好,否则,你就没办法正常的访问网络。下面,我们就来看看域名服务器的具体工作方式吧。

5. 域名服务器

上述的域名体系是抽象的,具体实现域名系统则是使用分布在各地的域名服务器。从理论上讲,可以让每一级的域名都有一个相对应的域名服务器,使所有域名服务器构成相对应的“域名服务器树”的机构。但是这样做会使得域名服务器的数量太多,使域名系统的运行效率下降。因此 DNS 就采用分区的办法来解决这个问题。

一个服务器所负责管辖的(或有权限的)范围叫作区。各单位根据具体情况来划分自己管辖范围的区。但在一个区中的所有结点必须是能够连通的。每一个区设置相应的权限域名服务器,用来保存该区中的所有主机的域名到 IP 地址的映射。

域名服务器的树状结构如图 3-5 所示。图中的每一个域名服务器都能够进行域名到 IP 地址的解析。当某个域名服务器不能进行域名到 IP 地址的转换时,它就设法找 Internet 上别的域名服务器进行解析。

从图 3-5 可以看出,Internet 上的域名服务器也是按照层次安排的。每一个域名服务器都只对域名体系中的一部分进行管辖。根据域名服务器所起的作用,可以把域名服务器划分为以下 4 种不同的类型。

1) 根域名服务器

根域名服务器是最高层次的域名服务器,也是最重要的域名服务器。所有的根域名服务器都知道所有的顶级域名服务器的域名和 IP 地址。不管是哪一个本地域名服务器,若要

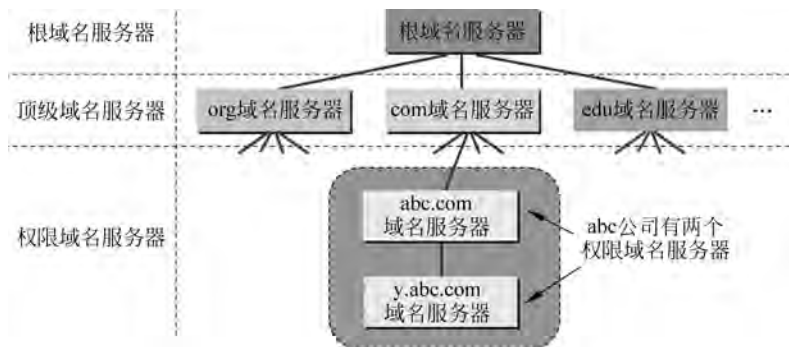


图 3-5 域名服务器的树状结构

对 Internet 上的任何一个域名进行解析,只要自己无法解析,就需要求助于根域名服务器。假定所有的根域名服务器都瘫痪了,那么整个的 DNS 系统就无法工作。

在 Internet 上共有 13 个不同 IP 地址的根域名服务器,名字是用一个英文字母命名的,从 a 到 m。这些根域名服务器的域名分别是 a. rootservers. net, ..., m. rootservers. net。但是,这些根域名服务器并不是简单的由 13 个机器所组成的(因为仅仅依靠 13 个机器,不可能为全世界的 Internet 用户提供令人满意的服务),而是 13 套装置。全球在全世界的不同地点安装了几百个根域名服务器机器。为了提供更可靠的服务,在每一个地点的域名服务器还可以由多台机器组成(为了安全起见,有些根域名服务器的具体地点还是保密的)。世界上大部分域名服务器都能就近找到一个根域名服务器。

必须指出的是,目前根域名服务器的分布仍然是很不合理的。在北美地区,平均每 370 万个网民就可以分摊到一个根域名服务器,而在亚洲地区,平均超过 2000 万个网民才分摊到一个,这样就会使亚洲地区的上网速度明显低于北美地区。

需要注意的是,在许多情况下,根域名服务器并不直接把待查询的域名直接转换成 IP 地址(根域名服务器也根本没有存放这样的信息),而是告诉本地域名服务器下一步应当找哪一个顶级域名服务器进行查询。

2) 顶级域名服务器

这些域名服务器负责管理在该顶级域名服务器注册的所有的二级域名。当收到 DNS 查询请求时,就给出相应的回答(可能是最后的结果,也可能是下一步应当找的域名服务器的 IP 地址)。

3) 权限域名服务器

这就是负责一个区的域名服务器。当一个权限域名服务器还不能给出最后的查询回答时,就会告诉发出查询请求的 DNS 客户,下一步应当找哪一个权限域名服务器。

4) 本地域名服务器

本地域名服务器并不属于图 3-5 所示的域名服务器层次结构,但它对域名系统非常重要。当一个主机发出 DNS 查询请求时,这个查询请求报文就发送给本地域名服务器。由此可以看出本地域名服务器的重要性。

为了提高域名服务器的可靠性,域名服务器都把数据复制到几个域名服务器来保存,其中的一个是主域名服务器,其他的是辅助域名服务器。当主域名服务器出现故障时,辅助域

名服务器可以保证 DNS 的查询工作不会中断。主域名服务器定期把数据复制到辅助域名服务器中,而更改数据只在主域名服务器中进行。

3.3.3 网络协议与 TCP/IP

【情景故事 3-9】 小红:对了,我还经常听说一个网络术语叫作协议,这又是什么意思呢?

小明:这个容易理解,俗话说“无规矩不成方圆”,任何事物都要有自己的规则,否则就会造成混乱。就好比两个人打电话,你说你的我说我的,大家同时说话,相互打断对方,那样说话就会变成吵架。两台主机之间的网络通信也是一样,如何建立连接、如何发送和接收信息、如何进行状态反馈,这些都需要在统一规则的约束下进行,而这个规则就是网络协议。

1. 协议

在计算机网络中要做到有条不紊地交换数据,就必须遵守一些事先约定好的规则。这些规则明确规定了所交换的数据的格式以及有关的同步问题。这些为网络中的数据交换而建立的规则、标准或约定称为网络协议(Network Protocol),简称为协议。网络协议主要由以下三个要素组成。

- (1) 语法,规定了数据与控制信息的结构或格式,包括数据出现的顺序;
- (2) 语义,规定了各种控制信息的意思,说明通信双方该怎么做;
- (3) 时序,也称为同步,规定了事件实现的顺序。

简单来说,就像中国和法国的两家企业的老总一起开会,语法就是大家都能理解的语言的语法(假定这种语言是英语);语义就是使用的英语单词和语句的意思;时序就是两位老总的秘书事先商量好的:谁先说、谁后说,先讨论什么内容、后讨论什么内容,语速是快还是慢等。

由此可见,网络协议是计算机网络不可缺少的组成部分。协议通常有两种不同的形式:一种是使用便于人来阅读和理解的文字描述;另一种是使用让计算机能够理解的程序代码。这两种不同形式的协议,都必须能够对网络上信息交换的过程做出精确的解释。

【情景故事 3-10】 小红:我懂了,看来网络协议是十分重要的,那 Internet 使用什么协议呢?它的内容又是什么呢?

小明:实际上,Internet 使用的不是一个协议,而是一套完整的协议系统。你想啊,像 Internet 这么复杂的计算机网络,它的规则想要用一两条简单的协议是很难描述清楚的,现在在互联网主要使用的是 TCP/IP 体系,它拥有多层次的协议体系结构,包含诸多具体的网络协议,其中最重要的就是 IP 和 TCP。

小红:为什么要使用多层次的结构呢?有什么好处?

小明:好处很多。划分层次之后,可以在各个层次使用不同的协议完成不同的工作,比如,在网络层使用 IP 实现主机的寻址、在运输层使用 TCP 完成可靠的数据传输等。各个层次功能独立,可以降低整个系统的复杂度,提高灵活性,同时也便于维护和开发。

小红:那 TCP/IP 体系结构是什么样的?一共分为几层呢?

小明:下面我们就来具体看看 TCP/IP 体系结构的内容。

2. TCP/IP 体系结构

目前,在 Internet 以及众多的局域网中使用的网络协议体系结构都是 TCP/IP 模型。

TCP/IP 体系结构如表 3-3 所示。

表 3-3 TCP/IP 体系结构

层 次	功 能 描 述
应用层	定义了 TCP/IP 及主机应用程序与网络运输层服务之间的接口
运输层	提供主机之间的通信会话管理,定义传输数据时的服务级别和连接状态
网络层	将数据装入 IP 数据报;包括用于在主机间及经过网络转发数据报时所用的源地址和目标地址信息;实现 IP 数据报的路由和寻址
网络接口层	通过网络,实现数据的实际物理传输;包括直接与传输介质接触的硬件设备、如何将比特流转换为电信号等

其中,核心为网络层的 IP 与运输层的 TCP,具体如下:

IP: TCP/IP 体系中的网络层协议。IP 可以提高网络的可扩展性:一是解决网络互联问题,实现大规模、异构网络的互联互通;二是分割顶层网络应用和底层网络技术之间的耦合关系,以利于两者的独立发展。需要注意的是,IP 只为主机提供一种无连接、不可靠的、尽力而为的数据包传输服务。

TCP: 一种面向连接的、可靠的、基于字节流的传输层通信协议。互联网络与单个网络有很大的不同,因为互联网络的不同部分可能有截然不同的拓扑结构、带宽、延迟、数据包大小和其他参数。TCP 的设计目标是能够动态地适应互联网络的这些特性,而且具备面对各种故障时的健壮性。

【情景故事 3-11】 小红:对了,我以前上网时有时会提示连接错误“无法连接 2xx.xxx.xxx.xxx(xxx)”,这里前面的一串数字的意思我现在明白了就是 IP 地址,那么后面括号里的数字又是什么呢?它有什么作用?

小明:这个数字叫作协议端口号,也叫作端口,用来标识收发网络数据的应用程序。

小红:应用程序?

小明:没错,我们在使用网络获取数据时,实际是通过主机上的某个应用程序来请求数据,如使用 QQ 收发信息、使用浏览器听歌等。那么当来自网络的数据通过 IP 地址找到我们的主机后,这些数据应该发给哪个应用程序呢?如果没有标识的话就会产生混乱,比如我们一边听歌一边在 QQ 上和朋友聊天,当网络数据到达时,如何判断这些数据代表的是歌曲信息还是朋友发来的 QQ 消息呢?这需要通过 IP 地址后的端口号来确定。不同的端口号对应不同的应用程序,如 QQ 对应的端口号可能是 5000,而浏览器对应的端口号可能是 80。

3. 协议端口号

协议端口号(Protocol Port Number)通常简称为端口(Port),用于区分进行数据传输的应用程序或者服务。使用端口以后,只需要把要传送的报文交到目的主机的某一个合适的目的端口,剩下的工作就由 TCP 来完成。这种协议端口是软件端口,和路由器或交换机上的硬件端口是完全不同的概念。硬件端口是不同硬件设备进行交互的接口,而软件端口是应用层的各种协议进程与运输实体进行层间交互的一种地址。

在数据传输过程中,一般使用一个 16 位端口号来标识一个端口。端口号只具有本地意义,它只是为了标识本地计算机应用层中的某个进程在和运输层交互时的层间接口。在

Internet 的不同计算机中,相同的端口号是没有关联的。16 位端口号可以允许有 65 536(0~65 535)个不同的端口号。

两个计算机的进程要相互通信,不仅需要知道对方的 IP 地址(为了找到对方的计算机),而且还要知道对方的端口号(为了找到对方计算机中的应用进程)。这和人们寄信的过程类似。当给某人写信时,就必须知道他的通信地址(IP 地址),在信封上还要写明收件人的姓名(端口号)。Internet 上的计算机通信是采用客户/服务器方式。客户在发起通信请求时,必须先知道对方服务器的 IP 地址和端口号。因此端口号分为以下两大类。

1) 服务器端使用的端口号

这类端口号又分为两类,最重要的一类叫作熟知端口号(Well-known Port Number)或系统端口号,数值为 0~1023。这些数值可在 www.iana.org 查到。IANA 把这些端口号指派给了 TCP/IP 最重要的一些应用程序,让所有的用户都知道。当一种新的应用程序出现后,IANA 必须为它指派一个熟知端口号,否则 Internet 上的其他应用进程就无法和它进行通信。常用的熟知端口号如表 3-4 所示。

表 3-4 常用的熟知端口号

应用程序	FTP	Telnet	SMTP	DNS	POP3	TFTP	HTTP	SNMP	SNMP(trap)
熟知端口号	21	23	25	53	110	69	80	161	162

另一类叫作登记端口号,数值为 1024~49 151。这类端口号是为没有熟知端口号的应用程序使用的。使用这类端口号必须在 IANA 按照规定的手续登记,以防止重复。例如,腾讯 QQ 服务器使用的端口号是 8000,新的版本的 QQ 客户端使用的端口号是 5000。

2) 客户端使用的端口号

数值为 49 152~65 535。由于这类端口号仅在客户进程运行时才动态选择,因此又叫作短暂端口号。这类端口号是留给客户进程暂时使用的。当服务器进程接收到客户进程的报文时,就知道了客户进程所使用的端口号,因而可以把数据发送给客户进程。通信结束后,刚才已使用过的客户端端口号就不复存在,这个端口号就可以供其他客户进程使用。

3.3.4 Internet 的接入

【情景故事 3-12】 小红:哎呀,说远了,IP 地址和 ISP 的意思我都明白了,但到底要如何才能连接 ISP,然后上网呢?

小明:嘿嘿,说到感兴趣的话题就不小心说远了。回归正题,我们把用户通过 ISP 接入到 Internet 的这一部分叫作接入网。接入网的覆盖范围一般为几百米到几千米,因而被形象地称为“最后一公里”。目前的接入技术主要是宽带接入技术。

小红:“宽带”这个词我知道,上网就是办宽带嘛,不过到底是什么意思就不清楚了。

小明:宽带接入是指上、下行速率分别不低于 512Kb/s 和 2Mb/s 的接入技术,总之就是传输速度比较快。宽带的接入方式有很多,根据传输介质类型的不同,分为有线接入和无线接入两大类,有线方式包括 ADSL、HFC 和光纤接入等等,无线接入有……

小红:等等,你说慢点。什么是传输介质,什么是 ADSL,完全不明白呢。

小明:好的,那我们就来具体了解一下网络传输介质和不同的接入方法吧。

1. 传输介质

传输介质是网络中连接收发双方的物理通路,也是通信中实际传送信息的载体。传输介质通常分为有线传输介质(导向型介质)和无线传输介质(非导向型介质)。

1) 有线传输介质

(1) 双绞线。

它是最古老但又最常用的传输介质。双绞线由两根分别包有绝缘材料的铜线螺旋状地绞合在一起,芯线为软铜线,线径为 $0.4 \sim 1.4\text{mm}$ 。两线绞合的目的是减少相邻线对之间的电磁干扰。

双绞线可以用来传输模拟信号和数字信号。双绞线的通信距离一般为几千米到十几千米。距离太长时就要加放大器以便将衰减了的信号放大到合适的数值(对于模拟传输),或者加上中继器以便将失真了的数字信号进行整形(对于数字传输)。导线越粗,其通信距离就越远,但导线的价格也越高。在数字传输时,若传输速率为每秒几兆比特,则传输距离可达几千米。由于双绞线价格便宜且性能也不错,因此使用非常广泛。

双绞线分为非屏蔽双绞线(Unshielded Twisted Pair, UTP)和屏蔽双绞线(Shielded Twisted Pair, STP),如图 3-6 所示。

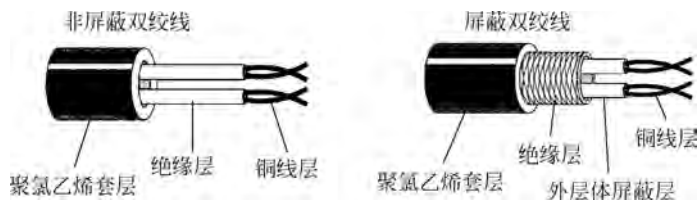


图 3-6 非屏蔽双绞线和屏蔽双绞线

(2) 同轴电缆。

同轴电缆(Coaxial Cable)由一根内导体铜质芯线外加绝缘层、密集网状编织外层体屏蔽层以及绝缘保护套层组成。同轴电缆结构示意图如图 3-7 所示。

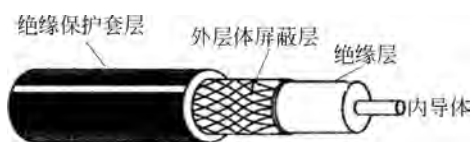


图 3-7 同轴电缆结构示意图

同轴电缆的特点是：高带宽及良好的噪声抑制性。同轴电缆的带宽取决于电缆长度, 1km 的电缆可以达到 $1 \sim 2\text{Gb/s}$ 的数据传输速率。通常,根据特性阻抗系数不同,分为 50Ω 同轴电缆和 75Ω 同轴电缆。

① 50Ω 同轴电缆：又称为基带同轴电缆或细缆,直接传送系带数字信号,传输速率最高可达 10Mb/s 。

② 75Ω 同轴电缆：又称为宽带同轴电缆、粗缆或 CATV 电缆。在计算机通信中,宽带系统是指采用了频分复用技术和模拟传输技术的同轴电缆网络。常用的 CATV 电缆在传输模拟信号时,频带高达 $300 \sim 450\text{MHz}$,距离可达 100km 。传输数字信号时,必须将其转换为模拟信号, 1b 占 $1 \sim 4\text{Hz}$ 的带宽。带宽为 300MHz 的 CATV 电缆可支持 150Mb/s ,通常传输一路电视节目占用 6MHz 的信道。

③ 光纤与光缆：光纤通信就是利用光导纤维(简称光纤)传递脉冲光来进行通信。由于可见光的频率非常高,约为 10^8MHz 的量级,因此一个光纤通信系统的传输带宽远远大

于目前其他各种传输媒介的带宽。

光纤通常有非常透明的石英玻璃拉成细丝,主要由纤芯和包层构成双层通信圆柱体。纤芯很细,直径只有 $8\sim 100\mu\text{m}$ 。当光线从高折射率的介质射向低折射率的介质时,折射角将大于入射角,当折射角足够大时,就会出现反射,即光线碰到包层时就会折射回纤芯。这个过程不断重复,光也就沿着光纤传输下去。光线在光纤中的折射如图 3-8 所示。

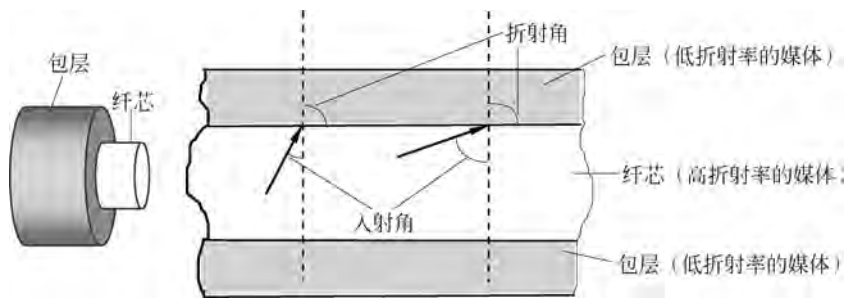


图 3-8 光线在光纤中的折射

光波在光纤中的传播过程如图 3-9 所示。

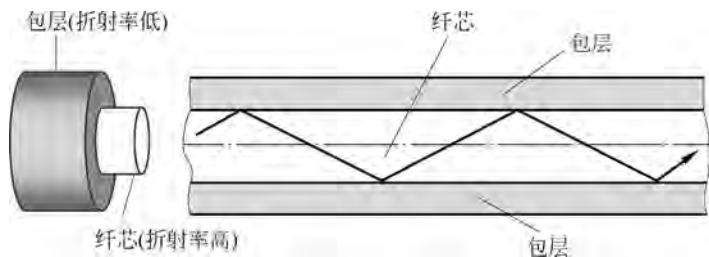


图 3-9 光波在光纤中的传播过程

光纤可分为多模光纤和单模光纤。使多条不同入射角的光线在一条光纤中传输,这种光纤就称为多模光纤(Multi-mode Fiber)。若光纤的直径减小到只有一个光的波长,则光纤就像一根波导那样,可使光线一直向前传播,而不会产生多次反射,这样的光纤称为单模光纤(Single-mode Fiber)。

由于光纤非常细,连包层合在一起的直径也不到 0.2mm 。因此,作为干线传输介质时,必须将光纤做成很结实的光缆。一根光缆少则只有一根光纤,多则可包括数十至数百根光纤,再加上加强芯和填充物就可以大大提高其机械强度。必要时还可放入远供电源线。最后加上包带层和外护套,就可以使抗拉强度达到几千克,完全可以满足工程施工的强度要求。五芯光缆的剖面示意图如图 3-10 所示。

光纤不仅具有通信容量非常大的优点,而且还具有传输损耗小、中继距离长、抗雷电和电磁干扰性能好、无串音干扰、保密性好、体积小、重量轻等特点。

2) 无线传输介质

对于有线传输介质来讲,若是通信线路要通过一些高山或岛屿,有时是很难施工的。即



图 3-10 五芯光缆的剖面示意图

使在城市中,敷设电缆也不是一件很容易的事情。当通信距离很远时,敷设电缆既昂贵又费时。但利用无线电波在自由空间的传播就可以较快地实现多种的通信。因此,就将自由空间称为无线传输介质(非导向型传输媒体)。无线传输介质包括:

(1) 短波。

短波通信主要是靠电离层的反射。通信频率范围为 $3\sim 30\text{MHz}$,通常称为高频(HF)段。由于电离层随季节、昼夜以及太阳黑子的活动而变化,所以通信质量并不稳定。

(2) 微波。

无线电微波通信在数据通信中占有重要地位。微波的频率范围为 $300\text{MHz}\sim 300\text{GHz}$ (波长为 $1\sim 10\text{cm}$),但主要是使用 $2\sim 40\text{GHz}$ 的频率范围。微波在空间中主要是直线传播,并且能够穿破电离层进入宇宙空间,因此它不像短波那样可以经电离层反射传播到地面上很远的地方。由于地球表面是一个曲面,所以一般在山顶建立微波中继站(简称“微波站”)。微波站的通信距离一般为 $30\sim 50\text{km}$,当微波天线高达 100m 时,通信距离可以达到 100km 。为实现远距离通信必须在一条微波通信信道的两个终端之间建立若干个中继站。中继站把前一站送来的信号经过放大后再发送到下一站,故称为“接力”。

微波通信的通信信道的容量很大,微波信号收到的干扰较小,传输质量较高。但是微波传播有时会受到恶劣气候的影响,其隐蔽性和保密性较差。

(3) 卫星。

常用的卫星通信方法是利用位于约 $36\,000\text{km}$ 高空的人造地球同步卫星作为中继器的一种特殊形式的微波接力通信。

和微波接力通信类似,卫星通信的频带很宽,通信容量很大,信号所受到的干扰也较小,通信比较稳定,并且卫星通信的通信费用与通信距离无关。卫星通信的另一特点就是具有较大的传播时延。

(4) 红外线通信和激光通信。

红外线通信和激光通信就是把要传输的信号分别转换为红外光信号和激光信号直接在自由空间沿直线进行传播,其特点是比微波通信具有更强的方向性,难以窃听、插入数据和进行干扰,但红外线和激光对雨雾等环境干扰特别敏感。

2. 宽带有线接入技术

1) ADSL

ADSL(Asymmetric Digital Subscriber Line,非对称数字用户线)技术是用数字技术对现有的模拟电话用户线进行改造,使它能够承载宽带数字业务。ADSL 的传输距离取决于数据率和用户线的线径(线径越细,衰减越大,传输距离越短)。ADSL 在用户线的两端各安装一个 ADSL 调制解调器,采用自适应调制技术使用户线能够传送尽可能高的数据率。ADSL 的上行信道带宽低于下行信道带宽。

2) 光纤同轴混合网

光纤同轴混合(Hybrid Fiber Coax,HFC)网是在目前覆盖面很广的有线电视网络基础上开发的一种居民宽带接入网,除可传送电视节目外,还能提供电话、数据和其他宽带交互型业务。

3) 光纤接入技术

光纤通信具有通信容量大、质量高、性能稳定、防电磁干扰、保密性强等优点。在干线通

信中,光纤扮演着重要角色。在接入网中,光纤接入也是发展的重点。

光纤接入方式可分为如下几种:FTTB(Fiber To The Building,光纤到大楼)、FTTC(Fiber To The Curb,光纤到路边)、FTTZ(Fiber To The Zone,光纤到小区)、FTTF(Fiber To The Floor,光纤到楼层)和 FTTH(Fiber To The Home,光纤入户)等。

4) 以太网接入

传统以太网不属于接入网范畴,而属于用户驻地网络领域,然而其应用领域正在向包括接入网在内的其他公用网领域扩展。对于企事业用户而言,以太网一直是最流行的组网技术,利用以太网作为接入手段的主要原因为:①以太网已有巨大的网络基础和长期的经验知识;②目前所有流行的操作系统和应用都与以太网兼容;③性价比高、可扩展性强、容易安装即高可靠性;④以太网接入方式与 IP 网相适应,同时以太网容量可以根据用户需要按实际情况设计以及升级。

【情景故事 3-13】 小红:这么多接入技术,每一个 ISP 都提供吗?那我应该如何选择呢?

小明:现在网络技术十分普及,不论是哪一个 ISP,大部分的地区都已经实现了光纤接入的方式,很多小区都已经能做到光纤入户,你只需要向一家 ISP 申请宽带接入业务,工作人员就会直接将光调制解调器(俗称光猫)与你家的光纤接口连接,再通过网线将计算机连接至光猫就可以上网了。

小红:那手机呢?手机要上网的话是不是也要进行 Internet 接入?

小明:当然了,现在手机、平板电脑等移动设备也十分普及了,所以越来越多的网络设备是通过无线移动接入技术接入 Internet 的,其中 4G 和 5G 技术是目前的主流及热点技术。

小红:对,经常听到 3G、4G 和 5G,就是不知道是什么意思。

小明:这些都是无线接入技术的名称,其中 G 就是英文 Generation,也就是代的意思,从 3G 到 5G,每一代都比前一代更加先进,速度更快,传输更稳定。当然,未来肯定还会有 6G(第六代)、7G(第七代)。

3. 4G 与 5G 接入

1) 4G 技术

4G 是 3G 的延伸,国际电信联盟(ITU)对 4G 的定义是静态传输速率达到 1Gb/s,用户在高速移动状态下可以达到 100Mb/s。与 3G 相比,4G 技术具有很多超越之处。其特点主要有高速率、以数字宽带技术为主、良好的兼容性、较强的灵活性、多类用户共存、多种业务的融合、高度自组织、自适应的网络等。

目前,4G 主要有以下两大标准。

(1) LTE 标准。

LTE 标准是 3G 的演进,其主要特点是在 20MHz 频谱带宽下能够提供下行 326Mb/s 与上行 86Mb/s 的峰值速率。与 3G 网络相比,4G 网络不仅大大提高了小区的容量,而且降低了网络延迟。

(2) WiMAX 标准。

WiMAX(Worldwide Interoperability for Microwave Access,全球微波互联接入)标准是一种基于 IEEE 802.16 标准的宽带无线接入城域网技术,是针对微波和毫微波频段提出的空中接口标准。采用 2G~11GHz 无须授权频段的宽带无线接入系统,其频道带宽可根

据需求在 1.5~20Mb/s 范围进行调整,所以 WiMAX 目前所使用的频谱可能比其他任何无线技术都要丰富。

2) 5G 技术

第五代移动通信技术(5th Generation Mobile Networks 或 5th Generation Wireless Systems、5th-Generation,简称 5G 或 5G 技术)是最新一代蜂窝移动通信技术,也是继 4G(LTE-A、WiMAX)、3G(UMTS、LTE)和 2G(GSM)系统之后的延伸。

5G 的性能目标是高数据速率、减少延迟、节省能源、降低成本、提高系统容量和大规模设备连接。Release-15 标准中的 5G 规范的第一阶段是为了适应早期的商业部署。Release-16 的第二阶段于 2020 年 4 月完成,作为 IMT—2020 技术的候选提交给 ITU。ITU IMT—2020 规范要求速度高达 20Gb/s,可以实现宽信道带宽和大容量 MIMO。

2019 年 10 月 31 日,我国三大运营商公布 5G 商用套餐,并于 11 月 1 日正式上线 5G 商用套餐。我国成为世界上最早、最大规模实现 5G 商用的国家。

5G 网络的主要优势在于,数据传输速率远远高于以前的蜂窝网络,最高可达 10Gb/s,比当前的有线互联网要快,比先前的 4G LTE 蜂窝网络快 100 倍。另一个优点是较低的网络延迟(更快的响应时间),低于 1ms,而 4G 为 30~70ms。

【情景故事 3-14】 小红:原来上网包含了这么多神奇的技术和内容啊。不过,我已经知道该怎么做了。小明,真是太感谢你了!如果不是你,我完全不知道该如何下手。

小明:不用客气,如果还有什么问题再随时联系我吧。再见。

3.4 Internet 的服务

3.4.1 Internet 的基础服务

【情景故事 3-15】 在小明的帮助下,小红家的宽带办好了,小红从此开始了在互联网世界遨游的日子。那么她每天都在网上干什么呢?最近,小红喜欢上了时下最流行的偶像团体 TFBOYS,尤其是其中的易烊千玺,小红简直被他迷得神魂颠倒,恨不得天天在网上浏览他的新闻和动态,耗费了好多的时间。小红也渐渐意识到这样消磨大好青春在网络上追逐偶像是不对的,但是就是控制不住自己,而且网上关于他的新闻一条接一条,好像无穷无尽一般,小红不禁好奇,互联网上这么多的信息和资源究竟是从哪里来的?她又是如何通过网络获取这些资源的呢?

其实,小红是通过信息浏览服务在 Web 上获取到这些资源的。下面就让我们了解 Internet 基础服务中信息浏览服务和 Web 的相关内容吧。

1. 信息浏览服务与 Web

信息浏览服务是目前应用最广泛的一种基本 Internet 应用。信息浏览服务是 Internet 资源共享的最好体现。

小红通过单击就可以浏览到的各种类型的信息,来自一个庞大的信息资源系统,这个系统称为环球信息网(World Wide Web,WWW),可以简称为 Web,中文翻译为“万维网”。它的正式定义为 WWW is a wide-area hypermedia information retrieval initiative to give

universal access to large universe of documents。Web 不是普通意义上的物理网络,而是一张附着在 Internet 上的覆盖全球的“信息网”,是一个大规模的、联机式的信息储藏所。严格来讲,Web 是一个技术系统,使用链接的方法能非常方便地从 Internet 上的一个站点访问另一个站点(也就是所谓的“链接到另一个站点”)。Web 提供的分布式服务如图 3-11 所示。



图 3-11 Web 提供的分布式服务

Web 有以下几方面的重要意义:

- Web 是一个支持多媒体的信息检索服务系统;
- Web 是一种基于超文本和超链接的信息处理技术;
- Web 是一种信息服务站点建设的规矩、规则和标准架构;
- Web 是 Internet 上提供共享信息资源站点的集合。

提供共享信息资源的站点称为“Web 网站”;承载资源信息内容的服务器称为“Web 服务器”。Web 服务器、超文本传输协议(HTTP)、浏览器是构成 Web 的三要素。在 Web 上资源信息使用专门的文档形式——网页(称为 Web 网页),记录、表示和存储;使用专门的语言——HTML,规范网页的设计制作;使用专门的技术——超链接技术,管理和组织众多的信息资源;使用专门的方法——统一资源定位器(URL),标识和寻址分布在 Internet 上的信息资源;使用专门的应用层协议——HTTP,实现数据信息的传送;信息检索服务采用 C/S 工作模式;在客户机上使用“浏览器”(如微软的 IE 浏览器)应用软件,实现信息浏览和检索。

(1) 超文本传输协议(Hyper Text Transfer Protocol, HTTP)是 Internet 上应用最为广泛的网络应用层协议,所有在客户端与 Web 服务器之间的信息传输都必须遵守这个协议。

HTTP 是一种请求/应答协议,定义了 Web 客户如何从 Web 服务器请求 Web 页面,以及 Web 服务器如何把用户需要的 Web 页面传送给客户。HTTP 还定义了 Web 页面的不同内容的现实顺序(如文本先于图形)等。当 Web 服务器对客户请求做出应答以后,连接便撤销,直到客户发送下一个请求才重新建立连接。HTTP 下的 WWW 浏览服务如图 3-12 所示。

HTTP 要求用户传递的信息只是请求方法和路径,整个协议规范比较简单,通信运行速度比较快,服务器规模可以比较小,并可以有效地处理大量请求。因此,HTTP 得到了广泛的应用,成为 Web 数据传输的标准协议。

(2) 网页(Web Page): 小红通过浏览器看到的信息组织形式就是网页,称为 Web 网页。网页通常使用超文本标记语言(Hyper Text Mark-up Language, HTML)设计制作,文件扩展名为 html、htm、asp、aspx、php、jsp 等。网页是构成网站的基本元素,是网站中的一

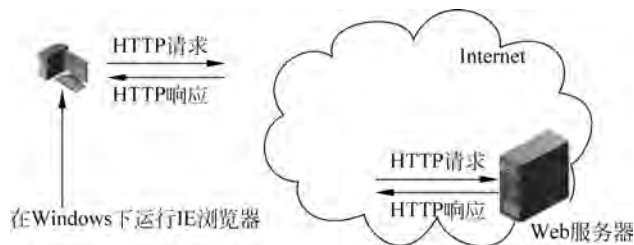


图 3-12 HTTP 下的 WWW 浏览服务

“页”，多个相关的网页合在一起，便组成了 Web 网站，如图 3-13 所示。从硬件角度来说，Web 网站是提供 Web 功能的服务程序。如果把 WWW 比喻成 Internet 的一个大型图书馆，Web 网站就像图书馆中的一本书，Web 网页就是书中的一页。

一个 Web 网站上存放着许许多多的网页，其中最受关注的是主页 (Home Page)。主页是一个 Web 网站的首页，从该页出发可以链接到本网站中的其他网页，也可以链接到其他网站。这样，就可以方便地接通 WWW 中任何一个 Internet 结点。主页文件名一般为 index、default。

Web 网页采用超文本的格式，可以包含文字、图像、声音、视频等信息，使 Web 网页的画面生动活泼，还可以含有指向其他 Web 网页或页面本身某特殊位置的超链接，这种包含超链接的文件称为超文本文件。

超链接首先是从一个网页指向一个目标的连接关系。这个目标可以是另一个网页，也可以是相同网页上的不同位置，还可以是一个图片、一个电子邮件地址、一个文件，甚至是一个应用程序。而在一个网页中用来超链接的对象 (称为超链接源) 可以是一段文本或者一个图片。当浏览者单击已经链接的文字或图片时，链接目标将显示在浏览器上，并且根据目标的类型来打开或运行。

超链接同时也是一种新型的、区别于线性方式的信息搜索技术。信息按线性方式搜索网页时，只能按网页的物理页码编号顺利地进行，这显然不适用于庞大的 Internet 搜索。超链接技术支持使用交叉的方式，借助于网页中包含着的“超链接源”，通过单击等方式，进行信息的快速搜索，大大提高了信息搜索的速度。超链接信息搜索过程如图 3-14 所示。



图 3-13 Web 网站

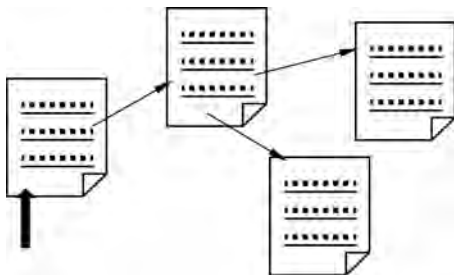


图 3-14 超链接信息搜索过程

(3) 统一资源定位器 (Uniform Resource Location, URL) 是 Web 中各种信息资源的编址标准，用于完整地描述 Internet 上网页和其他资源地址的命名和标识。Web 中信息资源是巨大的，同时也是具体的。每个承载着信息内容的网页都必须具有一个唯一的名称标识，

通常称为 URL 地址,俗称为“网址”,否则信息再丰富也不能实现便捷的访问浏览。

为保证信息资源命名的唯一性,URL 制定了统一的格式和规则。

URL 的一般使用格式如下:

```
scheme://host:port/path/filename
```

- scheme: 通信协议,指示该信息资源服务的协议类型。URL 中通信协议名称如表 3-5 所示。如果为 HTTP 服务,可省略“http://”。

表 3-5 URL 中通信协议名称

协议名称	功能
File	本地计算机上的文件资源
ftp	通过 FTP 访问的信息资源
Gopher	通过 Gopher 协议访问的信息资源
http	通过 HTTP 访问的信息资源
https	通过安全的 HTTP 访问的信息资源
mailto	资源为电子邮件地址,通过 SMTP 访问
News	通过 NNTP 访问的信息资源

- host: 主机名,只是提供信息服务的服务器域名或 IP 地址。
- port: 端口号,为可选项,只是提供信息服务所使用的端口号。如果使用的是 Internet 上信息服务的默认端口,则此项可以省略。比如,HTTP 服务的默认端口号为 80,如不重新设置改变端口,则端口号 80 就可以省略。
- path: 路径,指示资源文件在服务器中存放的路径。
- filename: 文件名,指示要访问的存放在服务器中指定路径下的资源文件的文件名。如果要访问的资源文件为网站的主页,则一般可省略此项。

例如:

```
http://wellan.znufe.edu.cn/2016/0316/c1668a34676/page.htm  
http://mail.163.com/js6/main.jsp?sid=GAMNNEpBFwLGvqTDdfBBAyVwECBlEgAp&df=email163#  
module=welcome.WelcomeModule%7C%7B%7D  
www.163.com
```

当通过搜索引擎搜索某指定信息资源时,显示搜索结果列表的网页地址中将包含搜索条件,其 URL 地址格式为 scheme://host:port/path? query# fragment。

- query: 搜索条件: 为可选项,用于给动态网页传递参数。可有多参数,用“&”符号隔开,每个参数的名和值用“=”符号隔开。
- fragment: 信息片段定位,为一个字符串,指定网络资源中的某个片段。例如,一个网页中有多个名词解释,可使用 fragment 直接定位到某一个。

万维网只是建立了 Web 网站的技术规则,但并没有制定对信息内容的约束。因此,任何用户只要遵循 Web 的技术规范,都可以把自己建立的信息发布到 Internet 上去。

【情景故事 3-16】 经过几天的冷静思考,小红意识到偶像虽然重要,但是自己不能因

为他而虚度光阴,荒废学业,她下决心将易烱千玺阳光向上的形象与不断努力的精神转化为自己学业进步的动力。考虑清楚之后,小红突发奇想,打算给易烱千玺写一封信,告诉自己的想法和决心。有了想法,小红打算先在网上查一查他家的地址,可该从何找起呢?小红陷入了思考。

Web 是一个大规模的、联机式的信息储藏所。如果已经知道所需信息的存放网点,那么只要在浏览器的地址框内输入该网点的 URL,按 Enter 键就可以进入该网点。但是,小红并不知道存有易烱千玺地址信息的网点 URL,这时便可以使用 Web 的搜索工具来对信息进行检索。下面我们就来看看互联网的海量信息之门——搜索引擎吧。

2. 搜索引擎——海量信息之门

在 Web 中用来进行搜索的工具叫作搜索引擎。搜索引擎的种类很多,大体上可以分为两大类,即全文检索搜索引擎和分类目录搜索引擎。

(1) 全文检索搜索引擎是一种纯技术型的检索工具,它的工作原理是通过搜索软件(Spider 程序)到 Internet 上的各网站收集信息,找到一个网站后可以从这个网站再链接到另一个网站,像蜘蛛爬行一样。然后,按照一定的规则建立一个很大的在线数据库供用户查询。用户在查询时只要输入关键词,就可以从已经建立的索引数据库上进行查询,但并不是实时地在 Internet 上检索到的信息,因此很可能有些查到的信息已经是过时的。建立这种索引数据库的网站必须定期对已经建立的数据库进行更新、维护。现在最著名的全文检索搜索引擎就是谷歌,它搜集的网页数量超过百亿,图片超过 10 亿张。中文搜索引擎中,最著名的是百度。

(2) 分类目录搜索引擎并不采集网站的任何信息,而是利用各网站在向搜索引擎提交的网站信息时填写的关键词和网站描述等信息,经过人工审核编辑后,如果认为符合网站登录的条件,则输入到分类目录的数据库中,供网上用户查询。

因此,分类目录搜索引擎也叫作分类网站搜索。分类目录的好处就是用户可根据网站设计好的目录有针对性地逐级查询所需要的信息,查询时不需要使用关键词,只需要按照分类(先找大类,再找下面的小类),因而查询的准确性较好。但分类目录查询的结果并不是具体的页面,而是被收录网站主页的 URL 地址,因而所得到的内容就比较有限。相比之下,全文检索可以检索出大量的信息,但缺点是查询结果不够准确,往往罗列了海量的信息,使用户无法迅速找到所需的信息。在分类目录搜索引擎中最著名的是雅虎。

从用户的角度来看,使用这两种不同的搜索引擎都能够实现自己查询信息的目的。但用户得到的信息的形式并不一样。全文检索搜索引擎往往可以直接检索到相关内容的网页,但分类目录搜索引擎一般只能检索到相关信息的网址。为了使用户能够更加方便地搜索到有用信息,目前许多网站往往同时具有全文检索搜索引擎和分类目录搜索引擎的功能。在 Internet 上搜索信息需要经验的积累,要多实践才能掌握从 Internet 获取信息的技巧。

目前,出现了垂直搜索引擎,它针对某一特点领域、特定人群或某一特点需求提供搜索服务。垂直搜索也是提供关键字进行检索的,但被放到了一个行业知识的上下文中,返回的结果更倾向于信息、消息、条目等。例如,对于寻找美食吃饭的人来说,他希望查找的是美食的具体信息(位置、价格、风格等),而不是有关于美食的一般性介绍或新闻、政策等。目前生活服务类是对垂直搜索需求最为旺盛的行业之一。还有一种是元搜索引擎,它把用户提交

的检索请求发送到多个独立的搜索引擎上去搜索,并把检索结果集中、统一处理,以统一的格式提供给用户,因此是搜索引擎上的搜索引擎。它的主要精力放在提高搜索速度、智能化处理搜索结果、个性化搜索功能的设置和用户检索界面的友好性上。元搜索引擎的查全率和查准率都比较高。

【情景故事 3-17】 小红耐心地在百度和雅虎两种不同类型的搜索引擎中尝试查找易烱千玺的地址信息,但遗憾的是都没能找到。小红遗憾地叹了一口气,正准备放弃时,不经意间在百度的搜索结果中发现了易烱千玺工作室的电子邮箱地址,于是她准备换一种方式,以电子邮件的形式给自己的偶像写信。

电子邮件也是 Internet 的一种基础服务,它可以把信息从一台计算机传送到另一台计算机。像传统的邮政服务系统一样,会给每个用户分配一个邮箱,电子邮件发送到收信人的邮箱中,等待收信人去阅读。下面我们就来具体了解一下电子邮件的相关内容吧。

3. 电子邮件

电子邮件(Electronic Mail, E-mail)是一种基于计算机网络的通信方式。它通过 Internet 与其他用户进行通信,往往在几秒或几分钟内就可以将电子邮件送达目的地,是一种快捷、简洁、高效和价廉的现代化通信手段。

在 Internet 中,电子邮件的传送、收发涉及一系列的协议,如 SMTP、MIME、POP3 和 IMAP 等。SMTP 用于在邮件服务器之间发送和接收邮件;MIME 用于对邮件及附件进行编码,实现在一封电子邮件中附加各种其他格式的文件;POP3 用于用户从该邮件服务器接收邮件;IMAP 提供了一个在远程服务器上管理邮件的手段。

SMTP(Simple Mail Transfer Protocol,简单邮件传送协议)是其中关键的一个协议。SMTP 是简单的基于文本的协议,其目标是可靠、高效地传送邮件。其协议端口号为 25。作为应用层的服务,SMTP 并不关心它下面采用的是哪一种传输服务,只要求有一条可以保证传送数据的通道。

邮件发送之前必须确定好邮件的发送者和接收者(即邮件地址)。用户并不是直接把邮件发给对方的邮件服务器,而是首先“联系”自己的邮件服务器,邮件服务器把邮件存放在缓冲队列当中。SMTP 客户通过定时扫描,发现队列中有待发送的邮件时,就和接收方的 SMTP 服务器建立 TCP 连接,并把邮件传送过去;如果在一定时间内邮件不能发送成功,则把邮件退还给发件人。

一个完整的电子邮件地址是由用户账号和电子邮件服务域名两部分组成,中间使用“@”相连,表示邮箱归属于以域名标记的电子邮件服务系统。如 hanmberg@163.com、liuqi1980@znufe.edu.cn 等。

用来收发电子邮件的软件工具很多,在功能和界面等方面各有特点,但它们都具有以下几个基本的功能,这些功能和人们日常生活中的邮政服务基本一致:

发送邮件:将编辑好的邮件连同邮件携带的附件一起发送到指定电子邮件地址。

阅读邮件:可以选择某一邮件,查看其内容。

存储邮件:可将邮件转存在一般文件中。

转发邮件:用户如果觉得邮件内容需要提供给他人,可在信件编辑结束后,根据有关提示转寄给其他用户。邮件服务及邮件传送如图 3-15 所示。

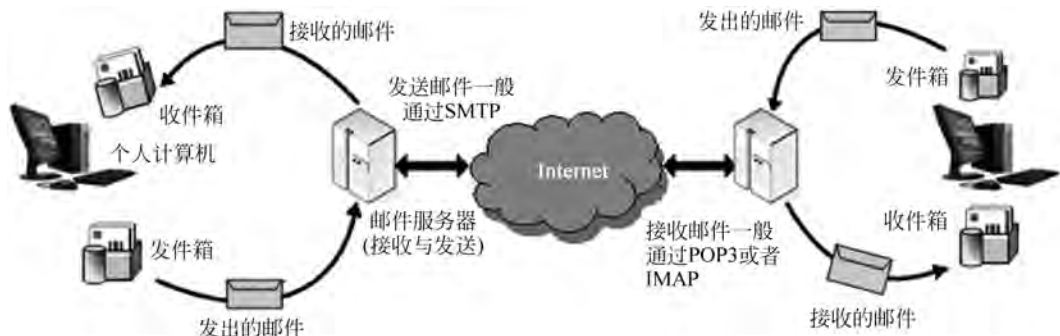


图 3-15 邮件服务及邮件传送

3.4.2 Internet 的现代信息服务

【情景故事 3-18】 结束自己的每日追星生活后,小红将主要精力放在了学业中,在期末考试中取得了很好的成绩,还拿到了一等奖学金。小红对自己的努力成果十分满意,决定买几件衣服犒劳一下自己,她一边进入淘宝网开始挑选商品,一边感叹现在网上购物的便利。暑假到了,小红的同学小露约她出国旅游,但小红还没有护照,也不知道该怎么办,小露告诉她现在办理护照可以在网上预约了,小红便按照小露的说明进入国家移民管理局政务服务平台,按照提示果然十分顺利地完成了预约操作,而且提醒了小红应该准备的材料,几天之后,小红拿着材料到出入境管理局办好了护照,连排队都不需要,真是方便极了。出发的日子快要到了,小红第一次出国旅游心里不免有些紧张,不知道该注意些什么,于是她打开了直播平台 App,关注了一个出国旅游达人,在看完他分享的一系列旅游经验和小贴士之后,小红顿时心里有了底,同时不免对这一次出行更加期待了。

网上购物、办业务、观看直播……现在人们生活已经离不开由 Internet 提供的各种新兴信息服务和产品,随着网络时代的发展,这些服务已经深入我们日常生活的每个角落,人们享受着互联网和大数据带来的便利。下面我们就通过几个具体的案例,了解一下这些 Internet 现代信息服务对人们生活的影响吧。

1. 直播电商——互联网+时代的新电商模式

2020 年 4 月 1 日,这是注定要被计入中国电商历史的一天。这天晚上 8 点整,“中国第一代网红”罗永浩在抖音正式开始他的首次带货直播,3 小时直播过程中的销售总额达到 1.1 亿元,整场直播观看总人数超过 4800 万,总销售件数逾 91 万。在销售的 22 款商品中,小米 10 系列售出 3800 多台,销售额超过 1600 万元;联想 Thinkplus 口红电源售出 3.6 万多个,迅速秒空下架;搜狗 AI 录音笔售出 1600 多台;米家声波电动牙刷售出超 10 万支;信良记小龙虾售出超 15 万份;小米巨能写中性笔售出 50 万支。这是罗永浩作为直播带货新人首秀交出的成绩。当晚同一时间,直播带货网红“一姐”薇娅也在直播间直播“卖火箭”,售价 4000 万元。淘宝数据显示,有超过 1970 万观众在线围观薇娅的这场直播。薇娅可以说是直播带货界的元老,早在 2018 年,薇娅就在“双 11”两小时销售 2.67 亿元;而 2019 年的“双 11”,薇娅只用了一天就完成了过去一年的销售额——27 亿元。薇娅淘宝直播“卖火箭”这一举动,被不少网友解读为与正在抖音首秀直播带货的罗永浩进行流量 PK。面对挑

畔,罗永浩在微博中罕见地温柔回应:中国消费品零售市场一年40万亿元,自己和其他几位带货网红目标人群并没有什么交集,“大家好好卖东西,卖好东西就是了”。

不经意间,直播电商正重新定义互联网+时代的电子商务模式。直播电商本质属于内容电商,是一种新的推销手段,直播为工具,电商为基础;通过流媒体的形式传播内容(促销内容、专业内容、娱乐内容),为电商带来流量,从而达到为电商销售的目的。直播电商最出名的代表当属“直播一姐”薇娅和“口红一哥”李佳琦。在这种主播大咖效应以及电商平台的助力之下,直播卖货逐渐成为一种潮流,淘宝、抖音、快手等都开始了直播卖货。因为在直播间,网友与网友之间、网友与主播之间具有社交互动,而直播间内动辄秒光的销售氛围,也如同肾上腺素,让部分消费者在观看的过程中产生冲动消费。而2020年初的新型冠状病毒肺炎疫情,对于直播卖货来说,如同“助推器”和“催化剂”,更是加速了直播卖货的发展。另外,直播电商商品的范围也不断扩大,从传统的衣服、化妆品、首饰,拓展到了土特产、房子等。淘宝直播是目前国内最大的直播平台。目前,年度用户超过4亿,场景覆盖全球73个国家的工厂、田间、档口、商场、街头、市场。去年一年,带动了400万人就业。2020年2月,淘宝新增商家数量达100万,淘宝直播新开播商家环比增幅达719%。与此同时,看直播购物也成为很多人的消费方式。2020年2月,在淘宝直播上“云生活”的用户增幅达153%,淘宝直播商家获得的订单总量平均每周都以20%的速度增长,成交金额比去年翻倍。如同网购给边远地区更多市场、机会一样,在网络基础设施、快递物流、网购理念、移动支付等多种配套设施逐步完善的背景下,直播带货也让更多三四线城市,农村甚至深山里,都有了接触更广阔市场的机会。从2019年开始,越来越多的村长、县长走进直播间,变身主播,吆喝着当地的土特产。2020年3月15日,淘宝直播举办了“春播月”活动,超过100位县长一起化身主播。其中,山东惠民县委副书记李宁波连续直播3h,带了30多款货,吸引了100多万网友关注,当场卖出约39000枚鸡蛋、7500根玉米、3000斤大蒜、26000斤香菇。

社交属性、营销话术、熟练程度、视频可见等多种因素叠加,让直播卖货越来越热。2019年,直播电商总成交总额超3000亿元,未来这一市场规模有望增长到万亿级。如今,直播电商完成由“人找货”向“货找人”的转变,依靠网红主播,紧密连接消费者与商家。电商平台改变了“以商品为媒介”的商业经营模式,开始转向“以人为媒介”。2020年3月30日,淘宝直播发布年度战略,未来一年将培养10万名月入过万的中小主播,在全国范围内打造100个“云市场”,帮助20万个线下门店通过直播拥抱新经济大潮。

目前,直播电商正处在天时地利的绝佳时机,部分线上需求的“爆发性增长”,也为市场生态长期变革埋下了伏笔。然而,在这个陡然加速的过程中,一些主播出售假冒伪劣产品、涉嫌虚假宣传、直播翻车的事件也随之浮现。根据中国消费者协会《直播电商购物消费者满意度在线调查报告》的数据显示:37.3%的受访消费者在直播购物中遇到过消费问题,消费者对主播夸大和虚假宣传、有不能说明商品特性的链接在直播间售卖等两点问题反馈较多。《电子商务法》虽然已经出台,但在落实过程中,还并未将直播平台特别是类似抖音、快手这样的综合类直播平台纳入其中。同时,由于行业的特殊性,平台在直播带货过程中到底扮演着什么样的角色?主播、销售商又该如何定位?这些都需要在未来更好地予以理清和规范。

2. 云端助力——电子政务让“信息惠民”落到实处

在湖北省武汉市“市民之家”,市民正在通过一站式服务办理户口迁移业务。据了解,由湖北电信承建的“市民之家”信息系统项目,采取主干核心万兆光纤+万兆交换机、千兆到桌面的

设计,使所有办理业务计算机“瘦终端”全部连接到统一的网上行政审批平台上,实现行政审批网络通、数据通、业务通。“市民之家”可同时容纳万名市民办理业务,并实现 24h 自助服务。

湖南益阳市市民孙先生身份证遗失需要补办,本以为是很麻烦的一件事情,没想到如今如此方便。他只是通过“益阳公安服务在线”网上办事项目,输入身份证号码,并用摄像头拍照保存,用网上银行付款,再留下邮寄地址,几分钟就完成了第二代身份证补办业务的网上申请,新身份证将直接被寄送到他的家中。

2015 年 12 月 23 日,武汉市推出集“市民卡”和“社会保障卡”功能于一体的“云端武汉·一卡通”市民卡,该卡以“社会保障卡”为“底本”,兼具政务、生活、金融服务三大功能,它既是持卡人享有政府服务、领取待遇的电子凭证,也可在公共交通、旅游景点、文化教育等公共领域作为银行卡使用。目前,已有 31 个部门(单位)的 172 项业务可凭该市民卡办理。同时,该市民卡可代替“武汉通”、购物可替代“支付宝”;去交管部门处理简易事故,或者借书、租车时,该市民卡相当于第二张“身份证”。

缴纳交通罚款或水电费指尖轻点便可完成、百姓办事找政府最多只需跑一次、新版“市民卡”可坐公交买东西……类似这样的便民举措如今已经数不胜数。一个个案例,一条条举措,折射出互联网时代电子政务发展的一系列历史性成就和变革。作为政府职能转变和管理创新的重要表现方式,电子政务已成为提高社会信息化水平的重要“抓手”。根据《国家电子政务“十二五”规划》,至“十二五”末,政务部门主要业务信息化应用水平将达到 85%,地市和县区分别平均达到 70%、50% 以上。如今,在云计算、物联网、移动互联网等新技术和新产业的助推下,电子政务在“信息惠民”中扮演着越来越重要的角色。从各地的实践来看,一个突出的变化就是直接面向老百姓的便民新应用正百花齐放地涌现出来,为他们提供“接地气”的服务。

“打破信息孤岛,打通数据壁垒”正成为中国电子政务工作的发展目标。而打破这一壁垒的武器正是以计算机网络为基础的云技术。建设服务政府管理和提供公共服务的统一政府云平台,被视为未来电子政务的“大脑”。以此为核心,政府各部门的公共信息资源得以共享和利用,同时还可以开展公共服务需求数据的跨部门分析。早在 2013 年,“上海失物招领服务平台”就通过云计算技术将包括派出所、轨道交通、铁路站点、出租车、公交车、机场、客运码头以及上海图书馆、上海博物馆等公共场所收到的失物信息整合到一起,并通过平台发布出来。截止到 2019 年,该平台已收到失物信息约 71 000 条,众多市民通过该平台找到失物。在山东新泰市,过去有 47 个部门分别建有独立的电子服务平台,其中 17 个部门有自己的专网,不能互联。但现在通过云计算,原来分散于市直各部门的服务器、网络设施、基础软件等信息化资源整合纳入市电子政务中心,打通的数据像血液一样通畅地流动在各个部门之间,为市民提供多样化的创新服务。

互联网时代电子政务的另一个发展方向是基于移动互联网的“移动政务”服务。在国外,移动政务正在蓬勃发展。以美国为例,美国有关政府部门将 2012 年称为移动政务元年,多个政府建立了移动政务平台,公众可以通过这一平台提出意见和想法,甚至可以参与投票表决。而在我国,移动政务也开始“崭露头角”,逐渐渗透至交通、公安、工商等多个部门。

在上海,越来越多的公交车站上,乘客只要扫一下二维码,就能知道本站所有公交车的路线和停靠站点,以及下一班公交车的到站时间。而在浦东的陆家嘴,借助智能停车系统,通过手机就能查询 1km 内的停车位。据估计,这项技术至少能“释放”出 5000 个车位。在

武汉,交警部门与阿里巴巴集团高德软件公司合作推出的“易行江城”App 可让武汉市民“一站式”享受行车、停车诱导和交管政务服务。在警民互动板块,市民可上报拥堵、故障、车祸、积水等 7 种交通事件,审核通过后,地图上会出现相应标注,并且在导航中进行规避。另外,市民可通过 App 实现一键报警。

此外,一些地区已经大力开展移动政务服务终端,形成“装在口袋里”的政务服务,越来越多的政府服务让群众办事跟网上购物一样省心、快捷。在多个城市,“市民主页”成为移动政务应用的统一入口,市民只需用手机访问和下载安装无线城市客户端后,进入“市民主页”页面,就可以使用各类服务。在北京,“市民主页”涵盖交通、教育、医疗、住房、旅游、政府办事等 12 大类便民服务项目,为市民提供包括车辆违规查询、网上挂号、缴纳水电费在内的涉及衣食住行等方面的共 200 余种政府政务信息以及服务应用。而在福建,市民则可将家里的水、电、燃气账号,社保、医保、公积金账号,驾照号,车牌号等与手机绑定,通过“市民主页”页面查询各种费用缴纳情况。在江苏,“互联网+政务”服务的“不见面审批”,让“不推一扇门,不见一个人,办成所有事”成为现实。江苏政务服务网自 2019 年 6 月上线以来,访问量突破 3.5 亿,移动客户端下载量达 2470 万,用户注册量 486 万,实名认证量 337 万,网络贯通全省 65 个省级部门、13 个市和 96 个县。

云技术与移动互联网的飞速发展,让百姓得到大量的便利,也将让中国的电子政务建设与落实拥抱数字新时代。

3. 数字战疫——互联网+助力抗击疫情

2020 年初,一场新型冠状病毒肺炎疫情席卷中国,面对来势汹汹的疫情,许多传统产业受到较大冲击,但互联网及其衍生的数字经济与产业模式却在此期间取得迅速发展,引领工作方式、生活方式、消费方式变革,表现出强大生命力,成为了缓解疫情影响、促进经济稳定增长的重要力量。疫情期间,以消费互联网为代表的数字经济满足了普通人居家隔离时的社交和购物需求,而产业互联网则通过助力各行各业,保障了整个社会的正常运转,也对社会长远发展产生了深远影响。

1) 互联网助力产业链协同

基于 5G、人工智能、大数据和云计算等新技术的产业互联网,已逐渐成为各行各业的新一代基础设施。通过推动数据的连接与共享,产业互联网进一步强化了各行各业的产业链协同效应,给中国制造创造了新的加速器,催生了一个又一个“速度纪录”。第一个“速度纪录”是“火神山”和“雷神山”两座方舱医院的建造。为了应对快速扩散的疫情、增强病患收治能力,中国政府迅速在武汉启动了两座医院的建造计划,创造了新的“速度奇迹”:拥有 1000 张床位的火神山医院和拥有 1600 张床位的雷神山医院,从建造到交付分别只用了 10 天和 12 天。支撑这一速度的是产业互联网带动的大规模协调能力。基于数字化平台和技术,设计单位在一天内迅速召集了数百位设计师,24h 内给出设计方案,60h 内便敲定了施工图纸。在施工现场,数千台大型机械的作业,在产业互联网平台上可以被实时监测与调度,确保施工准确推进。另一个速度纪录是大批“医疗用品制造商”的快速涌现。从 2020 年 1 月 1 日至 2 月 7 日,3000 多家中国企业新增了口罩、防护服和消毒液等医疗用品的生产业务,大幅提高了国内的防护物资产能,缓解了供应压力。企业能如此快速地实现跨界,得益于产业互联网为企业赋予的柔性生产能力:自上而下全面贯通的产业链以及模块化的流程设计,让企业在特殊时期能快速响应,调整生产线,实现跨界生产。

2) 互联网改变“宅家”生活

疫情时期的全民“宅家”生活,催生了一系列互联网新业态和新模式的飞速发展,网络办公、网络教学、网络医疗、网上消费、网上娱乐等互联网新模式取代了传统的工作、教学和生活消费模式。网络办公成为疫情期间主要的工作模式。2019年,我国远程办公人数约530万人。据统计,2020年春节复工之后,约有3亿用户使用远程办公,远程办公在复工30天内环比上升了663%,超过1800万家企业采用线上远程办公。据中国信通院抽样调查,90%信息消费企业采取“远程办公为主、驻地办公为辅”的开工模式。网络视频会议成为新的办公时尚。据统计,2019年,我国云视频会议增速达25%。疫情期间助推了云视频会议的井喷式增长,各级政府、学术机构、企业等大多采用网络视频会议的模式。用户注册量或者使用并发数量至少是2019年的10倍。中国移动开放云视讯软终端会议功能,湖北省累计开通软终端账号12万余个,协调数百家单位举办会议,参会人数达数万人。疫情以来,学校普遍延期开学,采用网络授课方式,在线教育规模迅猛增长。“老师变主播,家长成助教”,上网课已经成为疫情期间孩子们主要的学习方式。在疫情肆虐的日子,医院成为人们避之不及的场所,对于一般的小病小症,在线问诊成为重要的诊治手段。工信部的数据显示,目前全国191家公立医疗机构及近100家企业互联网医院针对疫情提供在线义诊。在武汉,受防控措施影响,市民除新冠肺炎以外的看病、复诊、拿药遭遇困难。在这一背景下,武汉医保局率先“破冰”,首次将平台型互联网医院——“微医”纳入医保支付。在上海,首家互联网医院牌照2020年2月26日“花落”徐汇区中心医院贯众互联网医院。它也是该市首家实现线上脱卡支付的公立互联网医院。上海市儿童医院互联网医院自2020年2月29日开展线上诊疗以来,1个月内已累计开展诊疗2170人次,开出网上处方579张。数字学习和数字娱乐快速增长。疫情期间在学习培训上的时间与花费均有较高提升,有21.1%的受访消费者表示自己首次尝试了在线培训或学习的行为。疫情期间,95%的网民增加了线上娱乐和服务的使用。爱奇艺、芒果TV、腾讯视频的网络视频会员数量分别环比增长了1079%、708%和319%。

3) 互联网加速金融业务转型升级

此次疫情无疑加速了金融业线上对线下业务的替代。疫情期间,互联网金融从业机构迅速通过门户网站、手机App、小程序等网络服务渠道,依托人工智能、大数据、生物识别、视频连线等科技手段,在线开展身份核验、资质审查、风险管理、签约办理、客服咨询等业务,及时为客户提供互联网支付、互联网借贷、互联网保险、互联网理财、互联网征信查询等多种金融服务,使人们担忧的人员聚集和面对面接触风险在这些业务领域得到了显著降低。证监会数据显示,疫情期间通过互联网渠道进行的证券交易超过95%。支付宝也推出在线理财业务,自2月4日后,支付宝的基金申购交易日均增长400%左右,用户大都购入股票类基金。目前支付宝已经与超过300家基金、银行、证券等金融机构展开合作。疫情期间,商业银行、第三方支付机构在大小额支付系统、网上支付跨行清算系统、支付机构网络支付清算平台等基础设施的有力支持下,为网络购物、生活缴费、资金汇划、捐赠捐助等提供安全、便捷的互联网支付,特别是移动支付展现了稳定高效的服务优势。包括新型互联网银行在内的各类银行机构以及消费金融、网络借贷等从业机构,依托人工智能、大数据等科技手段,精准识别疫情防控、复工复产和居民消费等方面的合理融资需求,及时为客户提供差异化、优惠性的互联网借贷服务。

4) 互联网+公共治理模式实现全民抗疫

疫情之下,在互联网技术助推下,各级国家机关和社会机构协同构建了新公共治理模式:政府机构积极引入合作伙伴,合力提升公共服务的效率和质量;社区和民众借助数字工具,也参与到社会的治理中。一个多元主体的公共治理模式渐渐成型,成为治理高度复杂的现代社会的范式。一方面,政府机构牵手数字化合作伙伴,高效指挥抗击疫情。在中国,许多地方政府部门都推出了数字政务服务,而疫情的爆发更是加速了政务数字化的进程。政府机构积极推动深度数字化,提升了物资调配、信息发布、民众意见采集等战役指挥功能的效率。截至2月8日,100多个政务类疫情服务小程序上线;政务类小程序用户在不到三周的时间内增长了近60%。数字化政务办理可以说已经成为各级政府机构的“基本配置”。另一方面,产业互联网赋能社区和民众,实现“全民抗疫”。在中国多个省市的社区里,居民借助数字平台组建了志愿者团队,提供消毒公共区域、配送社区物资等服务。此外,社区管理的数字化也极大地增强了社区的抗风险能力,各类数字化工具让每一个人都可以积极参与到疫情防控中来。除了前面提到的远程医疗等工具外,还有一个重要的数字化创新就是“健康码”。在“健康码”应用上,用户可以提交相关信息,如近期是否去过疫区、是否与确诊患者有过近距离接触等。“健康码”会通过不同的颜色,来显示用户近期与疫情相关的健康等级,从而让用户更好地配合相应的防控措施。

4. 短视频+直播——互联网流媒体服务的新方向

流媒体主要包括在线视频、短视频和直播,它是移动互联网时代的典型娱乐应用。随着互联网信息技术的升级,移动流量接入大幅增长,流媒体这类高耗流量应用的使用量大增,在线视频、短视频、网络直播合计使用时长占比达27%,仅次于社交网络居第二位。在流媒体的主要类型中,短视频和直播近几年发展迅速,为流媒体行业发展不断注入新的活力。而如今,短视频和直播行业已经开始相互渗透融合,并逐渐成为网络流媒体服务的新发展方向。从互联网领域各行业的发展来看,未来有可能形成类似阿里、携程、美团的综合型流媒体巨头。短视频与直播服务的融合,与两者当前的行业现状有着密切的关系。

2014年,游戏直播崛起,智能手机普及,迅速将视频直播的主战场从PC端转移到了移动端;2015—2016年,各大移动端的直播App蜂拥而至,2016年也成为直播元年,“千播大战”名噪一时。但经历短暂扩张期后,直播行业迅速进入洗牌期,在“放养”之后,国家正式对直播行业进行规范化管理。除了政策监管,直播行业本身过于庞大的竞争队伍也加速了内部的洗牌和重整。据《2016年度中国直播行业热度分析报告》不完全统计,2016年,中国直播平台有近200家,但下载量达千万的只有28家;根据公开数据显示,2017年,市场上存在的直播平台数量或已低于100家。过多的平台数量也使得直播用户群加速饱和。根据《2018—2019中国在线直播行业研究报告》的数据及预测,从2017年开始,中国在线直播用户的规模增速就在大幅下滑,预计2020年增速仅有4.59%,仅为2017年的1/6,行业已接近饱和。广为应用的“二八法则”在直播行业显得更极端。公开资料显示,2017年前五大直播平台占据了接近80%的市场份额,而在2017年中国游戏直播平台TOP1000的主播分布情况中,斗鱼和虎牙占据了市场超过70%的份额。2019年3月,王思聪投资并管理的熊猫直播平台宣布破产,也宣告直播行业迎来“洗牌2.0”时代。各大直播平台开始有了强烈的危机意识,要知道,熊猫直播在2018年第四季度的月活跃用户量达到了1095万,虽低于花椒、YY、虎牙和斗鱼,但也领先于企鹅、触手、映客等主流直播平台。

与直播行业类似,短视频行业的发展期也从2016年开始。早在2011年,快手就已经成立,随后陆续出现秒拍、美拍等产品,但短视频行业长期不温不火,直到2016年,抖音横空出世,一举奠定了“南抖音,北快手”格局下短视频行业的发展基础。与直播行业不同的是,短视频起源于传统网络流媒体行业,最初起始于土豆、优酷等网站推行的拍客短视频业务。并且,与直播的单打独斗相比,短视频行业更注重团队化竞争,腾讯、头条等旗下都拥有多款短视频产品。从发展历程上看,短视频行业正在走直播行业的老路。就在2016年,短视频行业度过井喷期后,不少产品出现内容违规、侵权等备受监管层关注的问题。面临同样严格的监管压力,直播和短视频却走上了两条截然相反的道路。中国互联网络信息中心发布的第43次《中国互联网络发展状况统计报告》显示,截至2018年12月,短视频用户规模达6.48亿,已高于网络视频用户规模的6.12亿,2018年下半年用户规模增长率达9.1%。相比之下,截至2018年12月,网络直播用户规模达3.97亿,较2017年年底减少2533万,用户使用率为47.9%,较2017年年底下降高达6.8%。除了直播行业本身的马太效应和高压监管外,造成这一现象的主要原因还有短视频行业的跨界抢占资源。抖音等短视频App纷纷开设直播功能,进一步加速直播行业的内部洗牌,更迫使头部直播企业向垂直化方向发展。比如,斗鱼涉足教育直播和泛娱乐化直播,虎牙以及企鹅电竞通过赛事直播签约职业战队等。

抖音、快手等短视频平台开启直播时代,直播行业也开始注重社交属性并加入了短视频内容。事实上,直播和短视频这两个行业正在相互融合。无论是短视频还直播,都可以看作是互联网时代下传统流媒体的精简版。不管是传统电视媒体还爱奇艺、腾讯音乐等流媒体,无非也是由视频/音频和直播两大部分组成的。

当前,整个流媒体行业正处于一个十字路口:究竟是各细分行业独立发展,还是由现有流媒体平台进行行业整合后,以全产业链形式发展?从整体来看,细分领域杀出的新兴行业和公司可能会有一段时期的高光表现;但从长远发展趋势来看,大行业与旗下包含的细分行业呈现出来的则是“分久必合,合久必分”。直播与短视频本就是传统媒体发展而来的新形式,从长远来看,在短视频和直播行业的风口过去之后,想要有更加长远的发展,或许就需要整个流媒体和传统媒体的介入整合。

3.5 网络数据获取

通过本章内容的学习,已经了解到利用Internet可以从Web中获取大量的信息。那么如何采集这些信息进而为大数据分析提供所需要的数据呢?最简单、直接的方法就是用Python的网络爬虫(Crawler)技术来解决。在本节中,将介绍网络爬虫的相关知识,引导读者使用Python语言构建网络爬虫并获取网络中的数据。

3.5.1 网络爬虫基础知识

1. 网络爬虫的概念与分类

人们通过浏览器来浏览网页,而网络爬虫是通过模仿浏览器来访问网页,它可根据某种规则自动获取所需要的网络信息。使用Python可以很方便地编写出爬虫程序,进行互联网数据的自动获取。爬虫又可分为通用爬虫和聚焦爬虫。其中,通用爬虫就是人们每天使

用的搜索引擎“抓取系统”的重要组成部分。其主要目的是将互联网上的网页下载到本地,形成一个对互联网已发布内容的镜像备份。通用爬虫会尽可能地把互联网上的所有的网页下载下来,放到本地服务器中形成备份,再对这些网页做进一步处理(如提取正文、去掉广告),最后提供一个用户检索接口。而聚焦爬虫是根据指定的需求抓取网络上指定的数据。例如:获取电影的名称和演员,而不是获取整张页面中所有的数据。聚焦爬虫会按照设定的规则,自动地抓取网页中的信息,并能沿着网页的相关链接在网络中采集资源,是一个功能很强的网页自动抓取程序。

目前网络爬虫已被广泛应用于搜集 Web 网页、文档、图片、音频、视频等资源。网络爬虫主要分成的 4 个步骤:①发送请求;②获取响应内容;③解析内容;④保存数据,如图 3-16 所示。

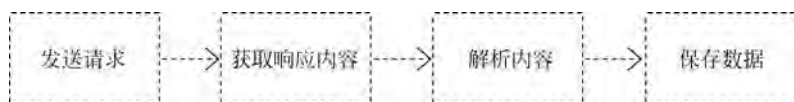


图 3-16 网络爬虫主要分成的 4 个步骤

2. HTML 简介

要从互联网中提取有用的数据,还需要了解用于创建的网页的标准标记语言:超文本标记语言(Hyper Text Markup Language,HTML)。HTML 也被称为网页源代码,它是通过一套标记标签来描述网页的一种语言,由标签和文本内容与属性构成。

HTML 标签是由大括号包围的关键词组成的,比如<html>。HTML 标签通常是成对出现的,比如和。标签对中的第一个标签是开始标签,第二个标签是结束标签。开始标签和结束标签也被称为开放标签和闭合标签。而开始标签和结束标签之间的文本被称为标签内容,如<p>这是标签内容</p>。

人们使用的网页浏览器(如 Chrome、Internet Explorer、搜狗、Safari 等)便是用于读取 HTML 文件,并将其内容显示出来的软件。如果需要用户查看 HTML 的源代码,以 Chrome 浏览器为例,可以通过在浏览器窗口右击,在弹出的快捷菜单中选择“查看网页源代码”命令,如图 3-17 所示。最终,查看网页源代码的结果通过 HTML 标签展示,如图 3-18 所示。

事实上,HTML 标签可转换为一棵 HTML 树,如图 3-19 所示,该树也被称为 DOM (Document Object Model)树,它是一种层次模型。DOM 树将网页中的各个元素都看作一个个对象,对象处于某个层次中,从而使网页中的元素也可以被计算机语言获取或者编辑。DOM 是以层次结构组织的结点或信息片断的集合,DOM 树把 HTML 文档呈现为带有元素、属性和文本的树结构。这个层次结构允许开发人员在树中导航以寻找特定信息。

3. requests 库的安装和使用

利用 Python 语言获取互联网上的 HTML 源代码首先需要安装第三方库——requests 库,requests 库的作用就是请求网站获取网页数据的,Python 的第三方库可以通过 pip 命令来安装。

在 cmd.exe 窗口中输入 pip 命令,如图 3-20 所示,如果返回 pip 命名的使用方法,说明 pip 命令可以正常使用。



图 3-17 选择“查看网页源代码”命令



图 3-18 查看网页源代码的结果通过 HTML 标签展示

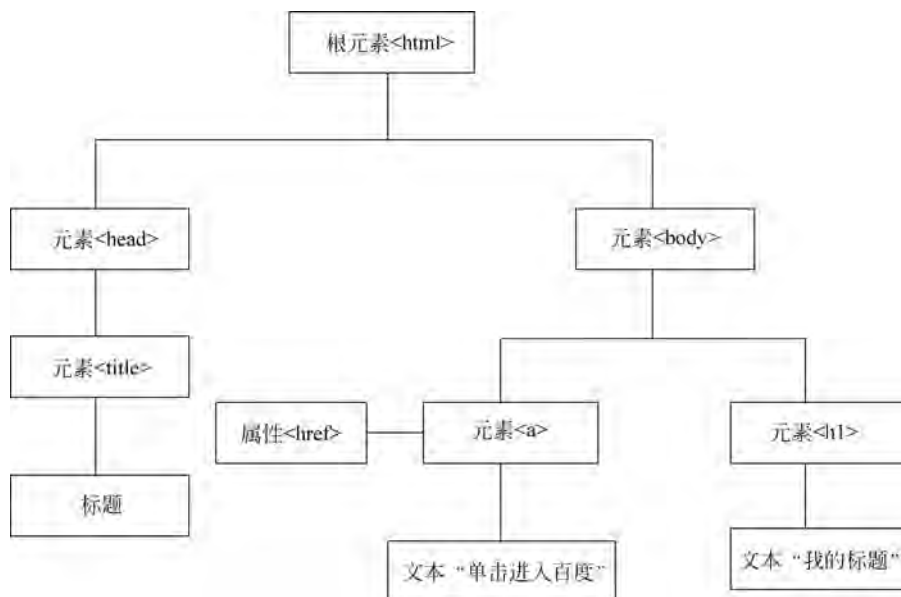


图 3-19 HTML 标签可转换为一棵 HTML 树



图 3-20 在 cmd.exe 窗口中输入 pip 命令

如果出现提示“‘pip’不是内部或外部命令,也不是可运行的程序”错误信息,则说明 Python 环境变量没有设置好,需要修复或者重新安装。

输入 `pip install requests` 完成 requests 库的安装,如图 3-21 所示。注意,安装的过程中需要计算机是联网状态,这样才能从网下获取第三方库。



图 3-21 输入 pip install requests 完成 requests 库的安装

在安装成功 requests 库之后,通过 requests 库输入网址获取网页内容。使用 requests 库获取网页内容最基本的方法是 `get()` 请求,例如获取访问百度主页的代码如下:

```
>>> import requests  
>>> url = "http://www.baidu.com"  
>>> res = requests.get(url)
```

通过 `requests.get()` 方法返回的是一个 response 对象,上面将其保存在 `res` 变量中,可以通过 `res` 变量来查看 response 对象的属性,其代码如下:

```
>>> res.status_code  
200  
>>> res.encoding  
'ISO-8859-1'  
>>> res.encoding = 'utf-8'
```

```
>>> res.text
<!DOCTYPE html>
<!-- STATUS OK --><html><head><meta http-equiv=content-type content=text/html;
charset=utf-8><meta http-equiv=X-UA-Compatible content=IE=Edge><meta content=
always name=referrer><link rel=stylesheet type=text/css href=http://s1.bdstatic.
com/r/www/cache/bdorz/baidu.min.css><title>百度一下,你就知道</title></head><body
link=#0000cc><div id=wrapper><div id=head><div class=head_wrapper><div class=
s_form><div class=s_form_wrapper><div id=lg><img hidefocus=true src=//www.
baidu.com/img/bd_logo1.png width=270 height=129></div><form id=form name=f action=//
www.baidu.com/s class=fm><input type=hidden name=bdorz_come value=1><input type=
hidden name=ie value=utf-8><input type=hidden name=f value=8><input type=
hidden name=rsv_bp value=1><input type=hidden name=rsv_idx value=1><input type=
hidden name=tn value=baidu><span class="bg_s_ipt_wr"><input id=kw name=wd class=
s_ipt value=maxlength=255 autocomplete=off autofocus></span><span class="bg_s_btn_
wr"><input type=submit id=su value=百度一下 class="bg_s_btn"></span></form></div>
</div><div id=u1><a href=http://news.baidu.com name=tj_trnews class=mnav>新闻
</a><a href=http://www.hao123.com name=tj_trhao123 class=mnav>hao123</a><a href=
http://map.baidu.com name=tj_trmap class=mnav>地图</a><a href=http://v.baidu.com
name=tj_trvideo class=mnav>视频</a><a href=http://tieba.baidu.com name=tj_trtieba
class=mnav>贴吧</a><noscript><a href=http://www.baidu.com/bdorz/login.gif?
login&tpl=mn&u=http%3A%2F%2Fwww.baidu.com%2f%3fbdorz_come%3d1 name=
tj_login class=lb>登录</a></noscript><script>document.write('<a href="http://
www.baidu.com/bdorz/login.gif?login&tpl=mn&u='+encodeURIComponent(window.location.
href+(window.location.search===' '? '?' : '&')+'bdorz_come=1')+'" name="tj_
login" class="lb">登录</a>');</script><a href=//www.baidu.com/more/ name=tj_
briiicon class=bri style="display: block;">更多产品</a></div></div></div><div
id=ftCon><div id=ftConw><p id=lh><a href=http://home.baidu.com>关于百度</a>
<a href=http://ir.baidu.com>About Baidu</a></p><p id=cp>&copy; 2017&nbsp;
Baidu&nbsp;<a href=http://www.baidu.com/duty/>使用百度前必读</a>&nbsp;<a href=
http://jianyi.baidu.com/ class=cp-feedback>意见反馈</a>&nbsp;<京 ICP 证 030173 号
&nbsp;<img src=//www.baidu.com/img/gs.gif></p></div></div></div></body>
</html>
```

在上述代码中, `res.status_code` 表示 `response` 对象的状态代码, 200 表示连接成功。`res.encoding` 表示 `response` 对象内容的编码方式。可以通过代码 `res.encoding = 'utf-8'`, 将原来的 ISO-8859-1 编码方式修改为 UTF-8 编码方式, 这样可以正常地显示中文字符。最终, 通过 `res.text` 用来显示通过爬虫获取到的网页内容。

通过 `requests.get()` 返回的 `response` 对象中, 包含一些常用的属性来表征请求响应后的结果。`response` 对象的常用属性及说明如表 3-6 所示。

表 3-6 response 对象的常用属性及说明

属 性	说 明
<code>res.status_code</code>	HTTP 请求的返回状态代码
<code>res.text</code>	URL 对应的网页内容
<code>res.encoding</code>	HTTP 响应内容的编码方式
<code>res.content</code>	HTTP 响应内容的二进制形式
<code>res.json()</code>	<code>requests</code> 中内置的 JSON 解码器
<code>res.raise_for_status()</code>	失败请求(非 200 响应)抛出异常

4. 网络爬虫需要遵守的协议

网络爬虫可从网络服务器爬取各种信息,其中可能存在涉及个人隐私或商业机密的内容,如果将不合适的内容提供给爬虫使用者,可能会给服务器管理者带来不必要的困扰与纠纷,因此需要 Robots 协议(Robots Exclusion Standard)来对网络爬虫进行规范。Robots 协议是网络爬虫排除标准,其作用为网站告知网络爬虫哪些页面可以抓取,哪些网页不可以抓取。该协议的内容存放于网站根目录下的 robots.txt 文件中。例如京东的 robots 文件网址为: <https://www.jd.com/robots.txt>,其内容如下:

```
User-agent: *
Disallow: /? *
Disallow: /pop/* .html
Disallow: /pinpai/* .html? *
User-agent: EtaoSpider
Disallow: /
User-agent: HuihuiSpider
Disallow: /
User-agent: GwdangSpider
Disallow: /
User-agent: WochachaSpider
Disallow: /
```

User-agent: * 中的 * 代表的所有的搜索引擎种类, * 是一个通配符,代表所有; Disallow: /? * 禁止访问所有以问号开头的链接; Disallow: /pop/* .html 代表禁止访问 pop 目录下以 html 结尾的网址; Disallow: /pinpai/* .html? * 代表禁止访问 pinpai 目录下 html 后面接 ? 号的网址。

当用 requests.get() 方法访问网页时, get() 方法后面还可以设置一个 headers 参数。当爬虫访问服务器时,有些服务器检查访问 headers 的 User-Agent 域,如果发现不是浏览器访问,服务器会拒绝访问。这时,可以手工加入 User-Agent 域,加入某种浏览器的请求头信息,其代码如下:

```
res = requests.get(url, headers = {'User-Agent': 'Mozilla/5.0'})
```

上述代码将模拟 Mozilla/5.0 浏览器访问相应的 URL 网页。

3.5.2 Python 网络爬虫实战

1. 基于正则表达式的数据获取

正则表达式(Regular Expression)是用于处理字符串的强大工具,通常被用来从文本中抽取符合某种规则的内容。Python 中使用正则表达式,只需要在程序前面加入 import re 即可。正则表达式拥有自己独特的语法,它们主要用于模式的匹配。其常见的语法如下:

- . 表示匹配任意字符,换行符\n除外。
- * 表示匹配前一个字符 0 次或无限次。
- ? 表示匹配前一个字符 0 次或 1 次。
- .* 表示贪心算法。

- . * ? 表示非贪心算法。
- () 表示括号内的数据作为结果返回。

在基于正则表达式的 Python 网络爬虫程序中,最常用的方法就是 `re.findall(pattern, string)`,该方法表示从 `string` 字符串中按照 `pattern` 这种模式去匹配内容。需要注意的是 `re.findall()` 方法的返回值为列表,之后还需要遍历这个列表来获取所需的内容。通过这种方法可以获取 `<li></li>` 标签中内容,即所有学院名称。其代码如下:

```
import re
s = "<html>
    <li>信息与安全工程学院</li>
    <li>新闻与传播学院</li>
    <li>法学院</li>
</html>"
results = re.findall('<li>(.*?)</li>', s)
for each in results:
    print(each)
```

其中, `s` 为所需要的处理的字符串,因为其中有换行,所以用三引号将其引起来, `re.findall(pattern, string)` 方法表示从 `string` 字符串中按照 `pattern` 这种模式去查找内容,方法的第一个参数代表匹配的模式,在上例中 `<li>(.*?)</li>` 表示按照左边是 `<li>`、右边是 `</li>` 的模式从 `s` 中抽取字符串。在上例中满足左边是 `<li>`、右边是 `</li>` 的恰好为所有学院的名称。

使用正则表达式的抽取方法主要就是观察需要抽取内容的左右两边,将需要抽取的中间内容用 `(.*?)` 代替即可,因为上例中的三个学院的名称都处于左边是 `<li>`、右边是 `</li>` 的标签之间,所以用模式 `<li>(.*?)</li>` 就可以抽取出它们的内容。

2. 基于 XPath 的数据获取

XPath 是 XML Path Language 的缩写,它是一种小型的查询语言,它既可以在 HTML 中查找信息,也可以通过元素和属性进行导航。第三方 XPath 库也需要先安装再使用,其安装方法为 `pip install lxml`。XPath 能够根据 HTML 语法建立解析树,进而高效解析其中的内容。XPath 使用路径表达式来选取 HTML 文档中的结点或者结点集。这些路径表达式和常规的计算机文件系统中看到的表达式非常相似。XPath 的简单调用方法代码如下:

```
from lxml import etree
selector = etree.HTML(网页源代码)
selector.xpath(路径表达式)
```

上面的源代码为网页的 HTML 源代码,路径表达式为 XPath 语法规则,XPath 主要有 6 种标签的使用方法:

- (1) `//`(双斜杠),定位根结点,会对全文进行扫描,在文档中选取所有符合条件的内容,以列表的形式返回。
- (2) `/`(单斜杠),寻找当前标签路径的下一层路径标签或者对当前路标签内容进行操作。
- (3) `/text()`,获取当前路径下的文本内容。
- (4) `/@xxxx`,提取当前路径下标签的属性值。

(5) . (点), 用来选取当前结点。

(6) .. (双点), 选取当前结点的父结点。

下面通过具体实例讲解通过 XPath 获取网页中的数据, 其代码如下:

```
from lxml import etree
html = "<html>
  <head>
    <title> Python 网络爬虫</title>
  </head>
  <body>
    <div class = 'university'>
      <a href = \"http://www.zuel.edu.cn/\">中南财经政法大学</a>
    </div>
    <div class = 'school'>
      <li><a href = \"http://xagx.zuel.edu.cn/\">工程学院</a></li>
      <li><a href = \"http://tsxy.zuel.edu.cn/\">数学学院</a></li>
      <li><a href = \"http://law.zuel.edu.cn/\">法学院</a></li>
    </div>
  </body>
</html> "
```

【例 3-1】 通过 XPath 获取标题数据“Python 网络爬虫”。

实现代码如下:

```
page = etree.HTML(html)
tags = page.xpath("//title/text()")
for tag in tags:
    print(tag) # 打印标题 Python 网络爬虫
```

在上面的代码中, 通过 //title 搜索根目录下的 < title > 标签, 通过 //title/text() 获取 < title > 标签里面的内容。上面代码的运行结果为:

```
Python 网络爬虫
```

【例 3-2】 通过 XPath 获取所有 < a > 标签里面的内容和超链接网址。

实现代码如下:

```
page = etree.HTML(html)
links = page.xpath("//a/@href")
for link in links:
    print(link) # 获取 href 的超链接
```

在上面的代码中, 通过 //a 搜索根目录下的 < a > 标签, 通过 //a/@href 获取所有的超链接信息。上面代码的运行结果为:

```
http://www.zuel.edu.cn/
http://xagx.zuel.edu.cn/
http://tsxy.zuel.edu.cn/
http://law.zuel.edu.cn/
```

【例 3-3】 通过 XPath 获取<a>标签中的学院内容,而不包括学校,可以先采用//div[@class='school']来获取学院所在的<div>标签,再获取<a>标签中的内容。

实现代码如下:

```
page = etree.HTML(html)
schools = page.xpath("//div[@class='school']/li/a/text()")
# 通过//div[@class='school'],搜索根目录下 class 属性值为 school 的<div>标签
for school in schools:
    print(school)
```

在上面的代码中,通过//div[@class='school']搜索根目录下 class 属性值为 school 的<div>标签,然后通过/li/a/text()搜索该<div>标签下标签下面的<a>标签内容,其运行结果为:

```
工程学院
数学学院
法学院
```

使用 XPath 的一个优势在于在多数浏览器中都有审查元素的功能,如 Chrome 浏览器右键快捷菜单中的“审查元素”命令,如图 3-22 所示。在右边选择所需的 HTML 代码,然后右击,在弹出的快捷菜单中选择 Copy XPath 命令复制对应的 XPath,如图 3-23 所示。通过 Copy XPath 可以直接将该元素的 XPath 代码复制到粘贴板中。



图 3-22 Chrome 浏览器右键快捷菜单中的“审查元素”命令



图 3-23 选择 Copy XPath 命令复制对应的 XPath

3. 使用 Python 操作 Excel 表格

Python 可以采用 Python 自带的 csv 库将数据写入 Excel 表格中。操作 Excel 表格时需要将 Excel 数据转换为一个二维列表,如图 3-24 所示。

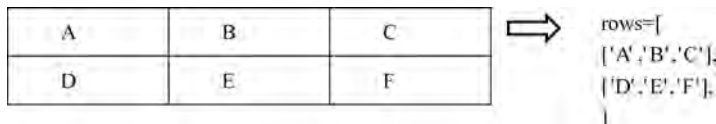


图 3-24 操作 Excel 表格时需要将 Excel 数据转换为一个二维列表

在上例中, `rows=[['A','B','C'], ['D','E','F']]` 实际上对应的就是 Excel 中第一行是 A,B,C,第二行是 D,E,F 的表格,在写入数据时,需要用 `open()` 方法建立一个 csv 文件并打开它,其代码如下:

```
file = open('test.csv', 'w', newline = '')
```

`open()` 函数的第一个参数 `test.csv` 是需要打开的文件名字,第二个参数 `w` 代表以写入的方式打开文件,第三个参数 `newline = ''` 表示在写入时行与行不需要空行。Python 写入 Excel 的代码如下:

```
import csv                                # 导入 csv 库
rows = [['A', 'B', 'C'], ['D', 'E', 'F']] # 定义二维列表
file = open('test.csv', 'w', newline = '') # 以写入的方式打开 test.csv 文件
f_csv = csv.writer(file)                 # 准备写入
f_csv.writerows(rows)                    # 写入数据
file.close()                             # 关闭文件
```

4. 基于正则表达式获取中南映像数据案例

下面介绍通过 Python 获取中南映像数据(网址为 <http://www.zuel.edu.cn/2019n/list.htm>),如图 3-25 所示。



图 3-25 Python 获取中南映像数据

源码: `target='_blank' title='25 万字，一份来自中南大的知识产权强国战略专家意见'><img`

↓

模式: `target='_blank' title='(.*)'><img`

↓

代码: `re.findall("target='_blank' title='(.*)'><img ", webpage)`

上例中的粗体部分“25 万字,一份来自中南大的知识产权强国战略专家意见”是需要从网页源代码中抽取的内容,只需要把粗体部分替换为(. * ?)就变成所需要的模式,最后在 Python 代码中使用 `re.findall("target='_blank' title='(. * ?)><img ", webpage)`,将会把 `webpage`(HTML 源码)中所有满足模式条件的字符串以列表的形式返回。

源代码: `<div class="wzly">供稿: 知识产权研究中心</div>`

↓

模式: `target='<div class="wzly">供稿: (.*)</div>'`

↓

代码: `re.findall("<div class=\"wzly\">供稿: (.*)</div>".webpage)`

上述粗体字部分“知识产权研究中心”是需要从网页源代码中抽取的内容，只需要把粗体部分替换为(.*?)就可以变成所需要的模式。在 Python 中使用 re.findall("<div class= \"wzly\">供稿：(.*?)</div>", webpage), 将会把 webpage(HTML 源代码)中所有满足模式条件的字符串以列表的形式返回。

```
import re
import requests
url = 'http://www.zuel.edu.cn/2019n/list.htm'
response = requests.get(url)
response.encoding = 'utf-8'
webpage = response.text
titles = re.findall("target = '_blank' title = '(. * ?)'><img", webpage) # 获取所有的标题信息
feeds = re.findall('<div class = "wzly">供稿: (. * ?)</div>', webpage) # 获取所有的供稿信息
print("所有标题: ")
for title in titles:
    print(title)
print("所有供稿单位: ")
for feed in feeds:
    print(feed)
```

最终程序的返回结果为所有的文章标题以及所有的供稿单位,返回结果内容如图 3-28 所示。

```

RESTART: C:/Users/Administrator/AppData/Local/Programs/Python/Python37/code/7基于正则表达式获取中南映像数据案例.py
所有标题:
25万字,一份来自中南大的知识产权强国战略专家意见'
【身边的榜样】指数经济创新团队党小组:党旗下的育才故事'
希贤好故事(23) 韩翼:把热情洒满新疆这片土地'
【身边的榜样群体】教学督导:做中南大教学质量的“提灯人”
石智雷:“学术研究的终点在于回馈学生”
“研中学”“做中学” 培养卓越会计人才'
“教书育人奖”获得者熊波:用诗句讲解大学数学的女老师'
“教书育人奖”获得者孙贤林:为学生义务补课19年'
“教书育人奖”获得者王淑红:坚守初心的“投资顾问”
“教书育人奖”获得者曹亮:教学能手 学生良师'
“教书育人奖”获得者黎江虹:先做益友 后为良师'
“教书育人奖”获得者刘惠好:把知识贡献给学生和社会'
“教书育人奖”获得者赵兴罗:将选修课变成“必修课”的财政史老师'
“教书育人奖”获得者卢现祥:新制度经济学的“传播者”
所有供稿单位:
知识产权研究中心
统计与数学学院
校报学通社
教学督导与评估中心
公共管理学院
会计学院
统计与数学学院
会计学院
工商管理学院
工商管理学院
法学院
金融学院

```

图 3-28 返回结果内容

5. 基于 XPath 获取中国银行股票数据案例

分析股票数据的获取对于金融分析非常有用,下面的案例将分析中国 2019 年第一季度的数据中国银行股票交易数据(网址为 http://quotes.money.163.com/trade/ljsysj_601988.html?year=2019&season=1),如图 3-29 所示。

中国银行 (601988) 历史交易数据										
2019 一季度 查询 下拉选择										
日期	开盘价	最高价	最低价	收盘价	涨跌幅	涨跌幅(%)	成交量(手)	成交金额(万元)	振幅(%)	换手率(%)
2019-03-29	3.71	3.79	3.71	3.77	0.06	1.62	1,791,175	67,257	2.16	0.09
2019-03-28	3.72	3.73	3.69	3.71	-0.02	-0.54	1,087,031	40,306	1.07	0.05
2019-03-27	3.73	3.75	3.72	3.73	0.01	0.27	1,058,660	39,510	0.81	0.05
2019-03-26	3.75	3.75	3.71	3.72	-0.01	-0.27	1,250,035	46,574	1.07	0.06
2019-03-25	3.78	3.79	3.73	3.73	-0.07	-1.84	1,880,907	70,767	1.58	0.09
2019-03-22	3.80	3.82	3.78	3.80	0.01	0.26	1,419,561	53,923	1.06	0.07
2019-03-21	3.80	3.82	3.79	3.79	0.00	0.00	1,713,724	65,226	0.79	0.08
2019-03-20	3.81	3.82	3.77	3.79	-0.02	-0.52	1,648,942	62,576	1.31	0.08
2019-03-19	3.84	3.85	3.80	3.81	-0.03	-0.78	1,601,599	61,221	1.30	0.08
2019-03-18	3.79	3.84	3.77	3.84	0.05	1.32	2,130,221	81,025	1.85	0.10
2019-03-15	3.79	3.82	3.78	3.79	0.00	0.00	1,518,035	57,629	1.06	0.07
2019-03-14	3.80	3.83	3.77	3.78	-0.01	-0.26	1,289,656	49,960	1.58	0.06
2019-03-13	3.83	3.84	3.78	3.80	-0.03	-0.78	1,581,609	60,540	1.57	0.08
2019-03-12	3.79	3.87	3.79	3.83	0.04	1.06	2,473,383	94,854	2.11	0.12
2019-03-11	3.74	3.81	3.74	3.79	0.04	1.07	1,981,058	74,496	1.87	0.09

图 3-29 分析中国 2019 年第一季度的数据中国银行股票交易数据

通过分析发现该网页的数据有两种不同的格式,其中表头的 HTML 格式如图 3-30 所示。



图 3-30 表头的 HTML 格式

对于以上 HTML 源代码采用 XPath 的路径表达式 `//tr[@class="dbrow"]/th/text()` 来解析里面的内容,其中 XPath 的前面部分 `//tr[@class="dbrow"]` 表示在整个网页中搜索 `class="dbrow"` 的 `<tr>` 标签, XPath 后面的 `/th/text()` 表示在该 `<tr>` 下面寻找所有的 `<th>` 标签里面的内容。

另外,表中数据的 HTML 格式如图 3-31 所示。

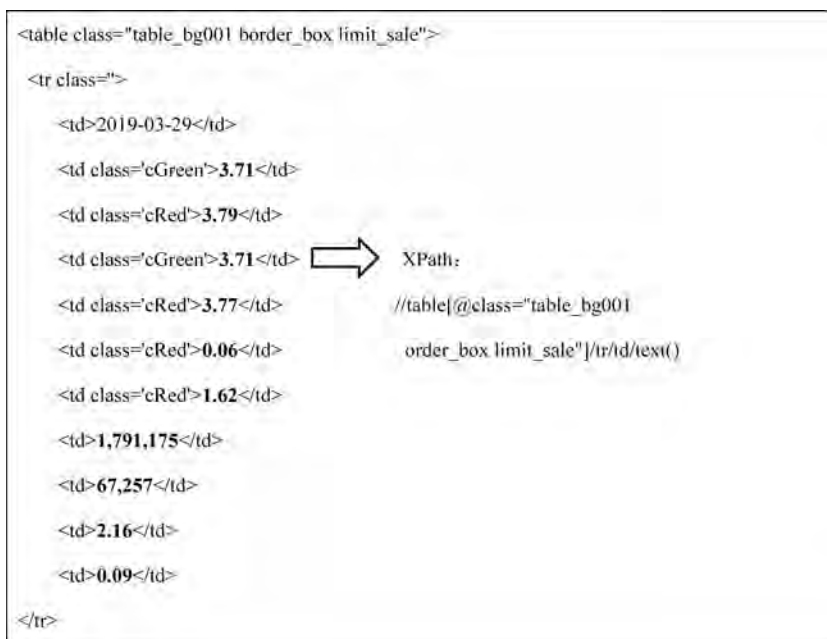


图 3-31 表中数据的 HTML 格式

其中,XPath 前面部分`//table[@class="table_bg001 border_box limit_sale"]`表示在整个网页中搜索 `class="table_bg001 border_box limit_sale"` 的`< table>`标签,XPath 后面的`/tr/td/text()`表示在该 `table` 里找到所有的`< tr>`下面的`< td>`数据,也就是所有的股票数据。

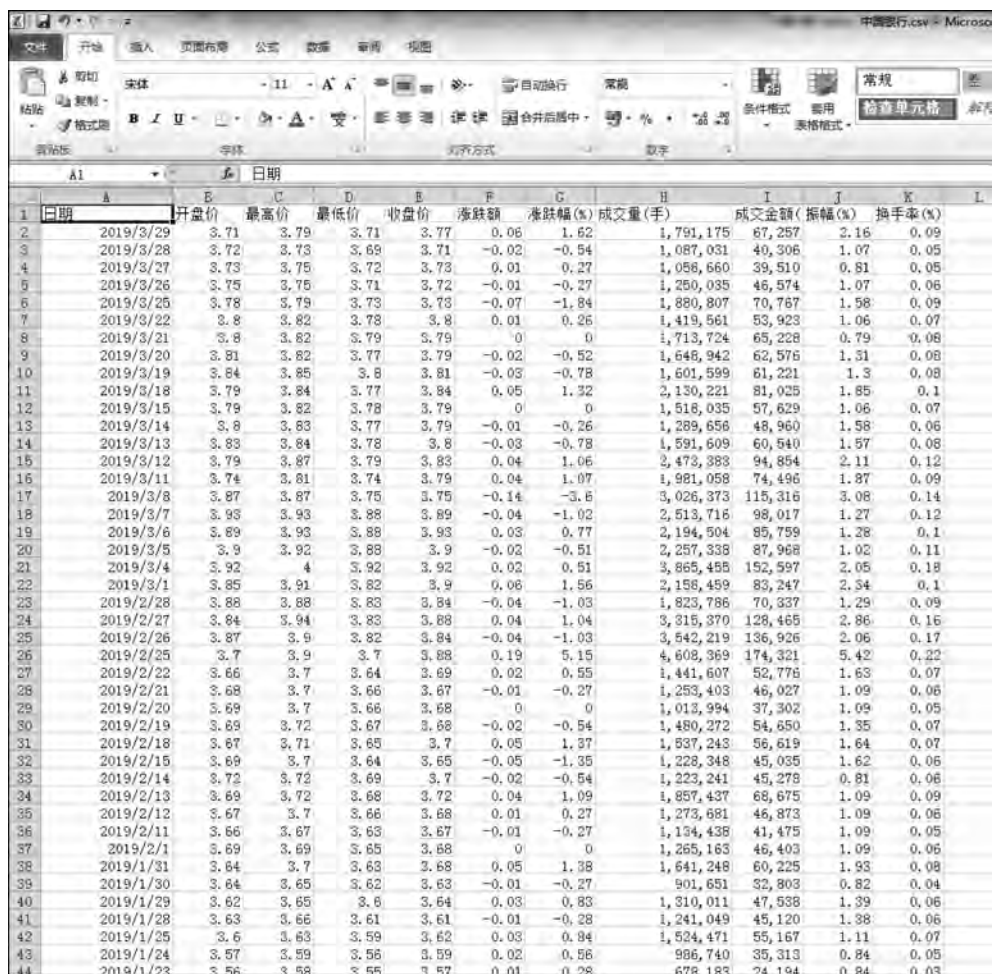
使用 Python 获取中国 2019 年第一季度的数据的完整代码如下:

```
import requests
import re
from lxml import etree
import csv
url = "http://quotes.money.163.com/trade/lsjysj_601988.html?year=2019&season=1"
# 设置待爬取的股票网址
colnum = 11 # 股票一共有 11 列属性
res = requests.get(url) # 访问股票网址
res.encoding = 'utf-8' # 设置为 utf-8 编码
content = res.text # 获取股票网页源代码
page = etree.HTML(content)
row = [] # 存储每一行数据
rows = [] # 按行存储所有数据
heads = page.xpath('//tr[@class="dbrow"]/th/text()') # 设置股票表头的 XPath
rows.append(heads) # 把表头加入到第一行
tds = page.xpath('//table[@class="table_bg001 border_box limit_sale"]/tr/td/text()') # 股票内容的 XPath
# i 用于计数,每 11 个为一组加入列表
i = 1
for td in tds:
    if(i % colnum != 0): # 如果不是 11 的倍数,就直接加入到 row 中
        row.append(td)
    else: # 每 11 个为一组存放到 rows 中
        row.append(td)
        rows.append(row)
        row = []
    i = i + 1
f = open('中国银行.csv', 'w', newline='') # 将数据写入 Excel 文件中
f_csv = csv.writer(f)
f_csv.writerows(rows)
f.close()
print(len(rows), "行数据写入完毕")
```

由于股票中的数据都是在`< td>`中,需要将其按照每 11 个为一组加入列表,于是采用下面的循环结构并结合 if 语句,代码如下:

```
i = 1 # i 用于计数
for td in tds:
    if(i % colnum != 0): # 如果不是 11 的倍数,就直接加入到 row 中
        row.append(td) # 把股票数据加入列表 row 中
    else: # 每 11 个为一组存放到 rows 中
        row.append(td) # 把股票数据加入列表 row 中
        rows.append(row) # 把一行数据加入到 rows 中
        row = [] # 清空 row 列表,用于下一次的数据的加载
    i = i + 1 # 计数器 i 增加 1
```

最终,程序将获取后的股票数据写入 Excel 文件中,如图 3-32 所示。



A1	日期	开盘价	最高价	最低价	收盘价	涨跌额	涨跌幅(%)	成交量(手)	成交额(元)	换手率(%)
1	2019/3/29	3.71	3.79	3.71	3.77	0.06	1.62	1,791,175	67,257	2.16
2	2019/3/28	3.72	3.73	3.69	3.71	-0.02	-0.54	1,087,031	40,306	1.07
3	2019/3/27	3.73	3.75	3.72	3.73	0.01	0.27	1,058,660	39,510	0.81
4	2019/3/26	3.75	3.75	3.71	3.72	-0.01	-0.27	1,250,035	46,574	1.07
5	2019/3/25	3.78	3.79	3.73	3.73	-0.07	-1.84	1,880,807	70,767	1.58
6	2019/3/22	3.8	3.82	3.78	3.8	0.01	0.26	1,419,561	53,923	1.06
7	2019/3/21	3.8	3.82	3.79	3.79	0	0	1,713,724	65,228	0.79
8	2019/3/20	3.81	3.82	3.77	3.79	-0.02	-0.52	1,646,942	62,575	1.31
9	2019/3/19	3.84	3.85	3.8	3.81	-0.03	-0.78	1,601,599	61,221	1.3
10	2019/3/18	3.79	3.84	3.77	3.84	0.05	1.32	2,130,221	81,025	1.85
11	2019/3/15	3.79	3.82	3.78	3.79	0	0	1,518,035	57,629	1.06
12	2019/3/14	3.8	3.83	3.77	3.79	-0.01	-0.26	1,289,656	48,960	1.58
13	2019/3/13	3.83	3.84	3.78	3.8	-0.03	-0.78	1,591,609	60,540	1.57
14	2019/3/12	3.79	3.87	3.79	3.83	0.04	1.06	2,473,383	94,854	2.11
15	2019/3/11	3.74	3.81	3.74	3.79	0.04	1.07	1,981,058	74,496	1.87
16	2019/3/8	3.87	3.87	3.75	3.75	-0.14	-3.6	3,026,373	115,316	3.08
17	2019/3/7	3.93	3.93	3.88	3.89	-0.04	-1.02	2,513,716	98,017	1.27
18	2019/3/6	3.89	3.93	3.88	3.93	0.03	0.77	2,194,504	85,759	1.28
19	2019/3/5	3.9	3.92	3.88	3.9	-0.02	-0.51	2,257,338	87,968	1.02
20	2019/3/4	3.92	4	3.92	3.92	0.02	0.51	3,865,455	152,597	2.05
21	2019/3/1	3.85	3.91	3.82	3.9	0.06	1.56	2,158,459	83,247	2.34
22	2019/2/28	3.88	3.88	3.83	3.84	-0.04	-1.03	1,823,786	70,337	1.29
23	2019/2/27	3.84	3.94	3.83	3.88	0.04	1.04	3,315,370	128,465	2.86
24	2019/2/26	3.87	3.9	3.82	3.84	-0.04	-1.03	3,542,219	136,926	2.06
25	2019/2/25	3.7	3.9	3.7	3.88	0.19	5.15	4,608,369	174,321	5.42
26	2019/2/22	3.66	3.7	3.64	3.69	0.02	0.55	1,441,607	52,776	1.63
27	2019/2/21	3.68	3.7	3.66	3.67	-0.01	-0.27	1,253,403	46,027	1.09
28	2019/2/20	3.69	3.7	3.66	3.68	0	0	1,013,994	37,302	1.09
29	2019/2/19	3.69	3.72	3.67	3.68	-0.02	-0.54	1,480,272	54,650	1.35
30	2019/2/18	3.67	3.71	3.65	3.7	0.05	1.37	1,537,243	56,619	1.64
31	2019/2/15	3.69	3.7	3.64	3.65	-0.05	-1.35	1,228,348	45,035	1.62
32	2019/2/14	3.72	3.72	3.69	3.7	-0.02	-0.54	1,223,241	45,279	0.81
33	2019/2/13	3.69	3.72	3.68	3.72	0.04	1.09	1,857,437	68,675	1.09
34	2019/2/12	3.67	3.7	3.66	3.68	0.01	0.27	1,273,681	46,873	1.09
35	2019/2/11	3.66	3.67	3.63	3.67	-0.01	-0.27	1,134,438	41,475	1.09
36	2019/2/1	3.69	3.69	3.65	3.68	0	0	1,265,163	46,403	1.09
37	2019/1/31	3.64	3.7	3.63	3.68	0.05	1.38	1,641,248	60,225	1.93
38	2019/1/30	3.64	3.65	3.62	3.63	-0.01	-0.27	901,651	32,803	0.82
39	2019/1/29	3.62	3.65	3.6	3.64	0.03	0.83	1,310,011	47,538	1.39
40	2019/1/28	3.63	3.66	3.61	3.61	-0.01	-0.28	1,241,049	45,120	1.38
41	2019/1/25	3.6	3.63	3.59	3.62	0.03	0.84	1,524,471	55,167	1.11
42	2019/1/24	3.57	3.59	3.56	3.59	0.02	0.56	986,740	35,313	0.84
43	2019/1/23	3.56	3.58	3.55	3.57	0.01	0.28	678,183	24,194	0.84

图 3-32 程序将获取后的股票数据写入 Excel 文件中

本章小结

计算机网络已经成为现代社会人们生活和工作中必不可少的基本工具和基本技术,了解计算机网络的基本原理、掌握计算机网络的基本技术以及应用,是一个现代人必备的素质。本章首先介绍互联网时代的发展历史,了解互联网技术是如何逐渐改变人们的生活的。接着介绍计算机网络的基础知识,学习何为计算机网络、如何构建计算机网络。随后介绍 Internet 的基础内容,通过案例引导读者了解 TCP/IP 结构及其核心内容,了解主机连接互联网的实现过程。接下来介绍 Internet 的相关服务,认识互联网究竟能够为人们带来哪些服务,并在人们生产生活的哪些方面发挥着重要的作用。最后,本章介绍了利用爬虫工具实现网络数据获取的方法,如此一来,就可以充分利用互联网中的海量信息和数据进行分析,从而获取更有价值的决策信息。

那么,通过网络爬虫工具获取分析所需的数据之后,接下来该如何处理?如何展示这些信息?在第2章中已经了解,大数据分析中涉及的数据类型繁多,常见的有结构化的数据和非结构化的数据。根据数据类型的不同,可以使用不同的方式进行处理和分析,而 Office 正是进行不同类型数据和信息的管理和处理的工具。在后续的章节中,将分别介绍利用 Office 2016 相关工具完成不同类型数据的处理和分析的方法。

思考题

1. 谈谈互联网时代对你身边不同年龄段人们生活的影响,比如对你、你的父母和你的爷爷奶奶生活的影响。
2. 你每天获取各类信息的主要手段都哪些?这些手段中哪些必须通过互联网?思考在获取的过程中可能涉及的本章学习过的知识。
3. 尝试自己动手编写一些 Python 爬虫程序,爬取自己喜欢的网站的数据。